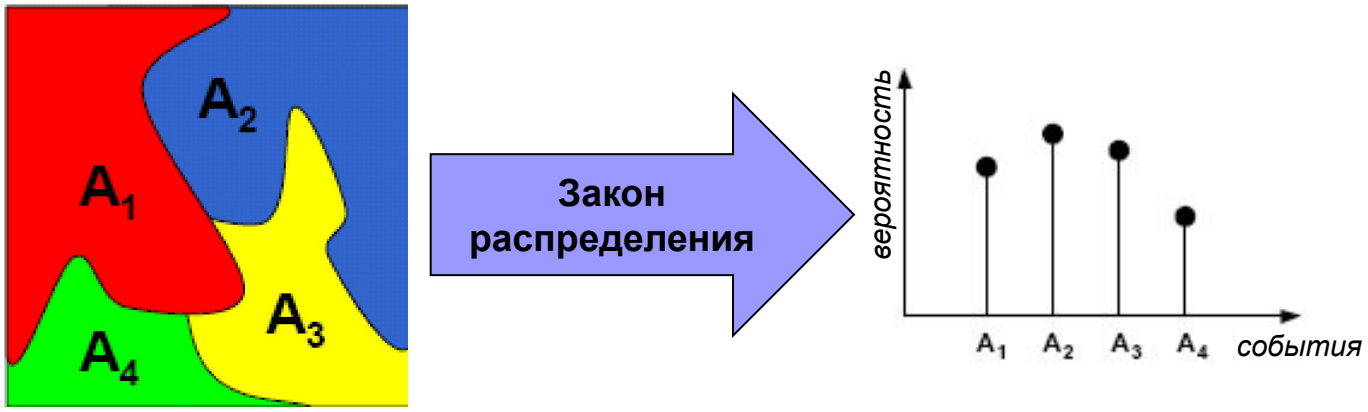
Байесовская классификация

Вероятность: основные понятия

- Определения (неформальные)
  - **Вероятность** – число, сопоставляемое событию и показывающее «насколько часто» будет происходить это событие при проведении случайного эксперимента
  - **Закон распределения вероятностей** в случайном эксперименте – правило, сопоставляющее вероятности событиям в эксперименте
  - **Пространство исходов S** случайного эксперимента – множество всех возможных итогов эксперимента



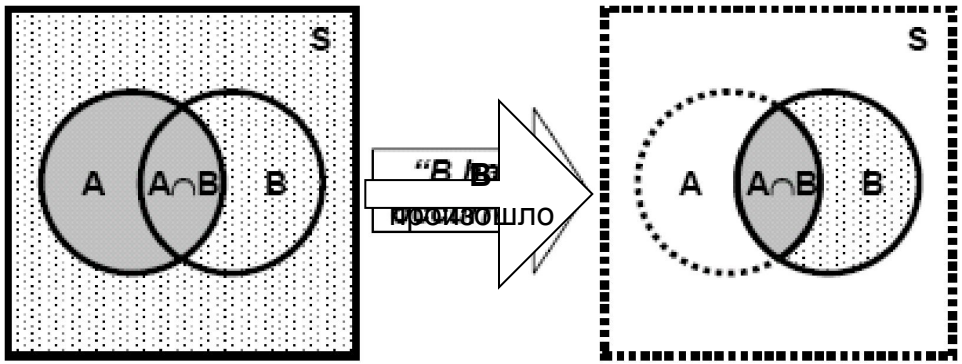
Вероятность: свойства

- Свойство 1: $P[A^C] = 1 - P[A]$
- Свойство 2: $P[A] \le 1$
- Свойство 3: $P[\varnothing] = 0$
- Свойство 4: Для $\{A_1, A_2, \ldots A_N\}$, если $\{A_i \cap A_j = \varnothing \ \ \forall i, j\}$,то $P[\bigcup_{k=1}^{N} A_k] = \sum_{k=1}^{N} P[A_k]$
- Свойство 5: $P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2]$
- Свойство 6: $P[\bigcup_{k=1}^{N} A_k] = \sum_{k=1}^{N} P[A_k] - \sum_{j<k}^{N} P[A_j \cap A_k] + \ldots + (-1)^{N+1} P[A_1 \cap A_2 \cap \ldots \cap A_N]$
- Свойство 7: Если $A_1 \subset A_2,$ то $P[A_1] \le P[A_2]$

Условная вероятность

- Если A и B – события, то вероятность события A при условии, что B уже произошло, определяется соотношением

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} \quad \text{для} \quad P[B] > 0$$

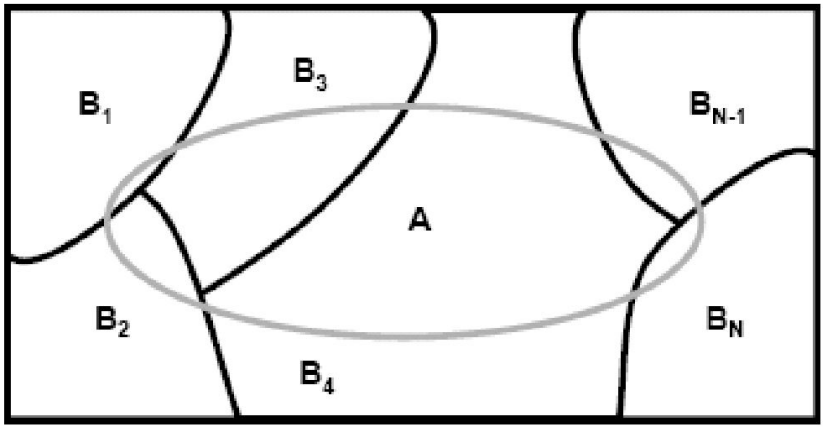


- Интерпретация
  - Новое обстоятельство «B произошло» имеет следующий эффект
    - Исходное вероятностное пространство S (весь квадрат) сужается до B (правый круг)
    - Событие A становится A∩B
  - Таким образом, P[B] просто нормирует вероятность событий, происходящих совместно с B

Формула полной вероятности

- Пусть B_1, B_2, …, B_N, – взаимоисключающие события, которые в объединении дают пространство исходов S (т.е разбиение S)
- Событие A может быть представлено как

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup ... \cup B_N) = (A \cap B_1) \cup (A \cap B_2) \cup ... (A \cap B_N)$$



- Поскольку B_1, B_2, …, B_N – взаимоисключающие, то

$$P[A] = P[A \cap B_1] + P[A \cap B_2] + ... + P[A \cap B_N]$$

- и, следовательно,

$$P[A] = P[A \mid B_1]P[B_1] + ... P[A \mid B_N]P[B_N] = \sum_{k=1}^{N} P[A \mid B_k]P[B_k]$$

Формула Байеса

- Пусть B_1, B_2, …, B_N дают разбиение S. Предположим, что произошло A. Какова вероятность B_j?
  - Из определения условной вероятности и формулы полной вероятности следует

$$P[B_j \mid A] = \frac{P[A \cap B_j]}{P[A]} = \frac{P[A \mid B_j] \cdot P[B_j]}{\sum_{k=1}^{N} P[A \mid B_k] \cdot P[B_k]}$$

Томас Байес (Thomas Bayes)
1702-1761

- Это соотношение известно как формула Байеса (правило Байеса, теорема Байеса) и является одним из самых полезных соотношений в теории вероятностей и статистике
  - В распознавании образов формула Байеса – один из фундаментальных результатов

Формула Байеса и статистическое распознавание образов

- Для решения задачи классификации формула Байеса может быть переписана следующим образом:

$$P[\omega_j \mid x] = \frac{P[x \mid \omega_j] \cdot P[\omega_j]}{\sum_{k=1}^{N} P[x \mid \omega_k] \cdot P[\omega_k]} = \frac{P[x \mid \omega_j] \cdot P[\omega_j]}{P[x]}$$

  - где ω_j – j-й класс, **x** – вектор характеристик образа
- Типичное решающее правило: выбрать класс ω_j, для которого $P[\omega_j \mid \mathbf{x}]$ – наибольшая
  - Интуитивно: выбираем класс, который наиболее ожидаемо даст **x**
- Каждый член в формуле Байеса имеет специальное название
  - $P[\omega_j]$ **априорная вероятность** (класса ω_j)
  - $P[\omega_j \mid \mathbf{x}]$ **апостериорная вероятность** (класса ω_j при условии, что в результате наблюдения получен **x**)
  - $P[\mathbf{x} \mid \omega_j]$ **правдоподобие** (условная вероятность появления наблюдения **x**, при условии, что наблюдается класс ω_j)
  - $P[\mathbf{x}]$ нормализующая константа, не влияющая на выбор решения

Простой пример

- Диагностическая задача: необходимо решить, болен ли пациент, основываясь на *несовершенном* тесте заболевания
  - Некоторые больные могут быть не распознаны (ложно-отрицательный результат теста)
  - Некоторые здоровые могут быть отнесены к больным (ложно-положительный результат теста)
- Терминология
  - Вероятность P[ОТР|ЗДОРОВ] истинно-отрицательного результата теста – **избирательность**
  - Вероятность P[ПОЛ|БОЛЕН] истинно-положительного результата теста – **чувствительность**

Простой пример

- Задача
  - Имеется выборка из 10000 человек, в которой больными являются по 1 человеку из каждых 100
  - Используется тест заболевания с избирательностью – 98% и чувствительностью – 90%
  - Допустим для Вас тест – положительный.
  - **Какова вероятность, что Вы больны?**
    - Решение 1: Заполнить таблицу сопряженности
    - Решение 2: Воспользоваться формулой Байеса

	Тест положителен	**Тест отрицателен**	**Итого**
Болен	Истинно-полож. P[ПОЛ|БОЛЕН]	Ложно-отриц. P[ОТР|БОЛЕН]	
Здоров	Ложно-полож. P[ПОЛ|ЗДОРОВ]	Ложно-отриц. P[ОТР|ЗДОРОВ]	
Итого			

Простой пример

- Задача
  - Имеется выборка из 10000 человек, в которой больными являются по 1 человеку из каждых 100
  - Используется тест заболевания с избирательностью – 98% и чувствительностью – 90%
  - Допустим для Вас тест – положительный.
  - **Какова вероятность, что Вы больны?**
    - **Решение 1: Заполнить таблицу сопряженности**
    - Решение 2: Воспользоваться формулой Байеса

	Тест положителен	**Тест отрицателен**	**Итого**
Болен	Истинно-полож. P[ПОЛ|БОЛЕН] **100x0.90**	Ложно-отриц. P[ОТР|БОЛЕН] **100x(1 – 0.90)**	**100**
Здоров	Ложно-полож. P[ПОЛ|ЗДОРОВ] **9900x(1 – 0.98)**	Ложно-отриц. P[ОТР|ЗДОРОВ] **9900x0.98**	**9900**
Итого	**288**	**9712**	**10000**

Простой пример

- Задача
  - Имеется выборка из 10000 человек, в которой больными являются по 1 человеку из каждых 100
  - Используется тест заболевания с избирательностью – 98% и чувствительностью – 90%
  - Допустим для Вас тест – положительный.
  - **Какова вероятность, что Вы больны?**
    - Решение 1: Заполнить таблицу сопряженности
    - **Решение 2: Воспользоваться формулой Байеса**

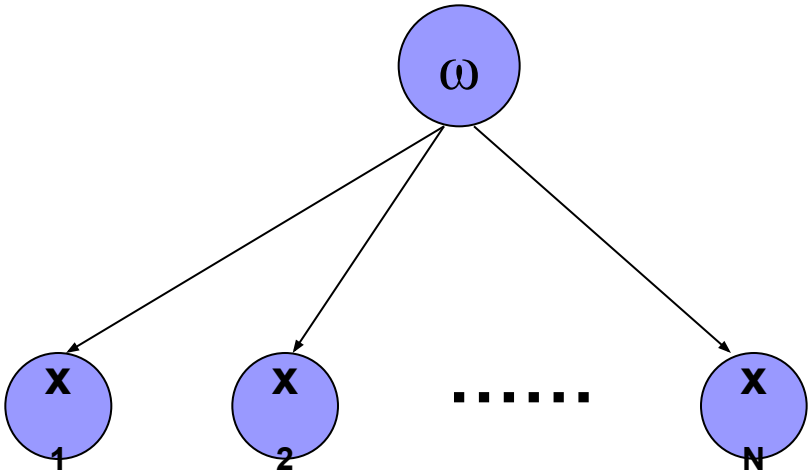
$$P[\text{БОЛЕН}|\text{ПОЛ}] = \frac{P[\text{ПОЛ}|\text{БОЛЕН}] \cdot P[\text{БОЛЕН}]}{P[\text{ПОЛ}]} =$$

$$\frac{P[\text{ПОЛ}|\text{БОЛЕН}] \cdot P[\text{БОЛЕН}]}{P[\text{ПОЛ}|\text{БОЛЕН}] \cdot P[\text{БОЛЕН}] + P[\text{ПОЛ}|\text{ЗДОРОВ}] \cdot P[\text{ЗДОРОВ}]} =$$

$$\frac{0.90 \cdot 0.01}{0.90 \cdot 0.01 + (1 - 0.98) \cdot 0.99} = 0.3125$$

Наивный байесовский классификатор

- Наивный байесовский классификатор – простой вероятностный классификатор, основанный
  - на формуле Байеса
  - на предположении **независимости** наблюдаемых признаков (что наивно)
- Графически
  - ω – класс (метка, номер класса), требуется оценить $P[\omega | x_1, x_2, \ldots, x_N]$
  - x_i – наблюдаемые признаки
  - все признаки независимы



Наивный байесовский классификатор

- Вероятностная модель наивного байесовского классификатора

$$\mathrm{P}[\omega \mid x_1,\ldots,x_N] = \frac{\mathrm{P}[\omega]\cdot \mathrm{P}[x_1,\ldots,x_N \mid \omega]}{\mathrm{P}[x_1,\ldots,x_N]}$$

- Строго говоря,

$$\mathrm{P}[\omega \mid x_1,\ldots,x_N] =$$
$$= \mathrm{P}[\omega]\cdot \mathrm{P}[x_1 \mid \omega]\cdot \mathrm{P}[x_2 \mid \omega, x_1]\cdot \mathrm{P}[x_3 \mid \omega, x_1, x_2]\cdot \mathrm{P}[x_4,\ldots,x_N \mid \omega, x_1, x_2 x_3]$$

- Но в виду условной независимости признаков $\mathrm{P}[x_i \mid \omega, x_j] = \mathrm{P}[x_i \mid \omega]$ и

$$\mathrm{P}[\omega \mid x_1,\ldots,x_N] = \mathrm{P}[\omega]\cdot \mathrm{P}[x_1 \mid \omega]\cdot \mathrm{P}[x_2 \mid \omega]\cdot\ldots\cdot \mathrm{P}[x_N \mid \omega]$$

- Тогда распределение условной вероятности класса ω есть

$$\mathrm{P}[\omega \mid x_1,\ldots,x_N] = \frac{1}{Z}\mathrm{P}[\omega]\prod_{i=1}^{N}\mathrm{P}[x_i \mid \omega]$$

где Z – константа, не зависящая от ω

Наивный байесовский классификатор

- Наивный байесовский классификатор – комбинация вероятностной модели и решающего правила «максимум апостериорной вероятности»

$$\text{classify}(x_1,\ldots,x_N) = \arg\max_j \mathrm{P}[\omega = \omega_j] \prod_{i=1}^{N} \mathrm{P}[x_i = f_i \mid \omega = \omega_j]$$

- Решающее правило
  - **Содержательно:** «Выбрать тот класс, который наиболее ‘вероятен’ для измерения **x**»
  - **Формально:** Вычислить апостериорную вероятность для каждого класса $\mathrm{P}[\omega_j|\mathbf{x}]$ и выбрать класс с ее наибольшим значением

- Для построения наивного байесовского классификатора
  - Оценить $\mathrm{P}[\omega=\omega_j]$, как долю объектов в обучающей выборке, для которых $\omega=\omega_j$
  - Оценить $\mathrm{P}[x_i=f_i \mid \omega=\omega_j]$, как долю объектов, для которых $\omega=\omega_j$ и при этом $x_i=f_i$
  - Для предсказания значения ω, найти то j, которое дает максимальное значение апостериорной вероятности

$$j^* = \arg\max_j \mathrm{P}[\omega = \omega_j] \prod_{i=1}^{N} \mathrm{P}[x_i = f_i \mid \omega = \omega_j]$$

Наивный байесовский классификатор

- **Задача**. Отнести текстовые документы к одному из предопределенных классов (спорт, политика, экономика,…)
- **Дано.** Имеется выборка из *N* текстов, разбитая на *K* групп
- **Решение**
  - Вычислить априорные вероятности классов
    - Посчитать количество текстов в каждом классе (группе) – N_j
    - Оценить априорную вероятность $P[\omega_j] = N_j / N,\ j=1,\ldots,K$
  - Вычислить условные вероятности появления слов в текстах разных классов
    - Пусть есть словарь из *M* слов: $x_1, x_2, \ldots x_M$
    - Вычислить c_{ij} – сколько раз слово x_i встретилось в текстах класса ω_j
    - Вычислить n_j, сколько слов из словаря встречается в текстах класса ω_j
    - Условные вероятности слов есть $P[x_i \mid \omega_j] = c_{ij} / n_j,\ j=1,\ldots,K$

Наивный байесовский классификатор

- **Задача**. Отнести текстовые документы к одному из предопределенных классов (спорт, политика, экономика,…)
- **Дано.** Имеется выборка из *N* текстов, разбитая на *K* групп
- **Решение**
  - Классифицировать новый текст *t*:
    - Вычислить признаки *t* (посчитать сколько раз входит слово x_i в *t*)
    - $P[\omega_j \mid t] = P[\omega_j \mid x_1, x_2, \ldots x_M] = P[\omega_j]\, P[x_1 \mid \omega_j]\, P[x_2 \mid \omega_j] \ldots P[x_M \mid \omega_j]$
    - Отнести *t* к классу ω_j с максимальной апостериорной вероятностью $P[\omega_j \mid t]$

Наивный байесовский классификатор

- **Задача**. Отнести текстовые документы к одному из предопределенных классов (спорт, политика, экономика,…)
- **Дано.** Имеется выборка из *N* текстов, разбитая на *K* групп
- **Решение**
  - Предобработка
    - Устранение пунктуации
    - Устранение чисел
    - Приведение регистра (все буквы - строчные)
    - Устранение слов короче 4 символов
  - Построение словаря и оценка встречаемости слов производится с помощью хэш-таблицы
  - Как выбрать ключевые слова (признаки)?
    - Взять *k* наиболее популярных слов в каждом классе
    - Взять *k* наиболее популярных слов в выборке
    - Объединить все эти слова. Вектор признаков – количества этих слов.

Наивный байесовский классификатор

- **Задача**. Отнести текстовые документы к одному из предопределенных классов (спорт, политика, экономика,…)
- **Дано.** Имеется выборка из *N* текстов, разбитая на *K* групп
- **Решение**
  - Прочие нюансы
    - Нулевые вероятности «убивают» байесовский классификатор
      - Если слово встречается только в одном классе, то условные вероятности такого слова «обнулятся»
      - Для предотвращения этого условные вероятности следует оценивать как $P[x_i \mid \omega_j] = \varepsilon / n_j$, где ε – малая настраиваемая константа
    - Все вероятности имеет смысл логарифмировать, чтобы избежать переполнения (особенно при вычислении длинного произведения)
    - Непрерывные признаки имеет смысл квантовать (дискретизировать)

Критерий отношения правдоподобия

- Пусть объект классифицируется на основании измерения (вектора характеристик) **x**

- Разумное решающее правило:
  - «Выбрать тот класс, который наиболее ‘вероятен’ для измерения **x**»
    - Более формально: вычислить апостериорную вероятность для каждого класса $P(\omega_j | \mathbf{x})$ и выбрать класс с ее наибольшим значением

Критерий отношения правдоподобия

- В случае 2-классовой задачи классификации:
  - *Если* $P(\omega_1 | \mathbf{x}) > P(\omega_2 | \mathbf{x})$, выбрать ω_1, *иначе* выбрать ω_2
  - Или, короче: $$P(\omega_1 | x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2 | x)$$
  - Используя формулу Байеса, получим $$\frac{P(x | \omega_1)P(\omega_1)}{P(x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(x | \omega_2)P(\omega_2)}{P(x)}$$
  - P(X) не влияет на решающее правило и может быть исключено $$\Lambda(x) = \frac{P(x | \omega_1)}{P(x | \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$
  - $\Lambda(x)$ называется **отношением правдоподобия**, а решающее правило известно как **критерий отношения правдоподобия** (КОП)

Критерий отношения правдоподобия: пример

- Используя критерий отношения правдоподобия, построить решающее правило для следующих условных плотностей классов (полагая априорные вероятности равными):

$$P(x\mid\omega_1)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-4)^2} \qquad P(x\mid\omega_2)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-10)^2}$$

- Решение
  - КОП для заданных плотностей:

$$\Lambda(x)=\frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-4)^2}}{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-10)^2}}\underset{\omega_2}{\overset{\omega_1}{\gtrless}}1$$

  - Упрощаем:

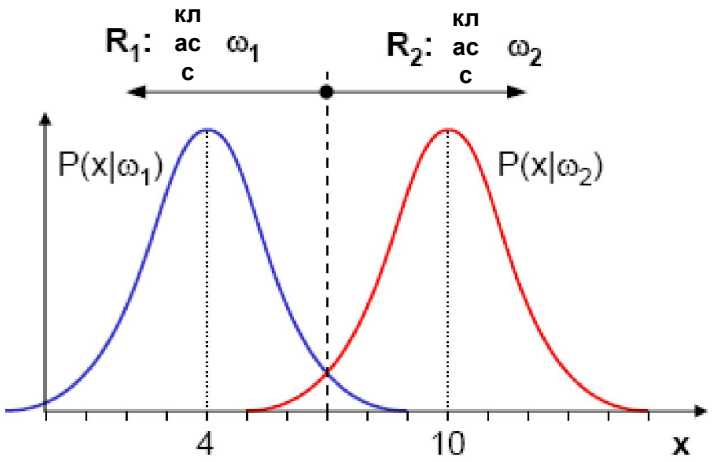
$$\Lambda(x)=\frac{e^{-\frac{1}{2}(x-4)^2}}{e^{-\frac{1}{2}(x-10)^2}}\underset{\omega_2}{\overset{\omega_1}{\gtrless}}1$$

  - Меняем знаки и логарифмируем:

$$(x-4)^2-(x-10)^2\underset{\omega_2}{\overset{\omega_1}{\lessgtr}}0$$

  - В итоге:

$$x\underset{\omega_2}{\overset{\omega_1}{\lessgtr}}7$$



Критерий отношения правдоподобия: пример

- Предыдущий пример понятен с интуитивной точки зрения, т.к. правдоподобия идентичны и отличаются только средними значениями

- Как изменится критерий отношения правдоподобия, если, например, априорные вероятности такие, что $P(\omega_1) = 2P(\omega_2)$?

Вероятность ошибки

- Мерой качества любого решающего правила может выступать с **вероятность ошибки** P[error]
- В соответствии с формулой полной вероятности может быть представлена как

$$P[error] = \sum_{i=1}^{C} P[error \mid \omega_i]P[\omega_i]$$

- Условная вероятность ошибки класса P[error | ω_i]:

$$P[error \mid \omega_i] = P[\text{выбран } \omega_j \mid \omega_i] = \int_{R_j} P(x \mid \omega_i)dx$$

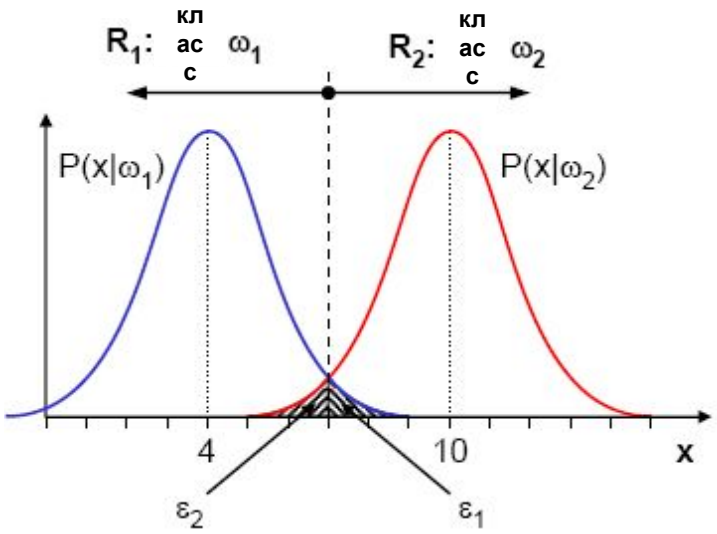
- Для 2-классовой задачи вероятность ошибки примет вид

$$P[error] = P[\omega_1]\underbrace{\int_{R_2} P(x \mid \omega_1)dx}_{\varepsilon_1} + P[\omega_2]\underbrace{\int_{R_1} P(x \mid \omega_2)dx}_{\varepsilon_2}$$

  - где ε_i – интеграл правдоподобия P(x | ω_i) по области R_j, соответствующей выбору ω_j

Вероятность ошибки

- Для решающего правила из рассмотренного примера интегралы ε_1 и ε_2 показаны на рисунке
  - Поскольку априорные вероятности полагались равными, то $P[error] = (\varepsilon_1 + \varepsilon_2)/2$



Вероятность ошибки

- Выясним, насколько хорош критерий отношения правдоподобия в смысле вероятности ошибки

  - Для этого удобно выразить P[error] в терминах апостериорной вероятности P[error | x]

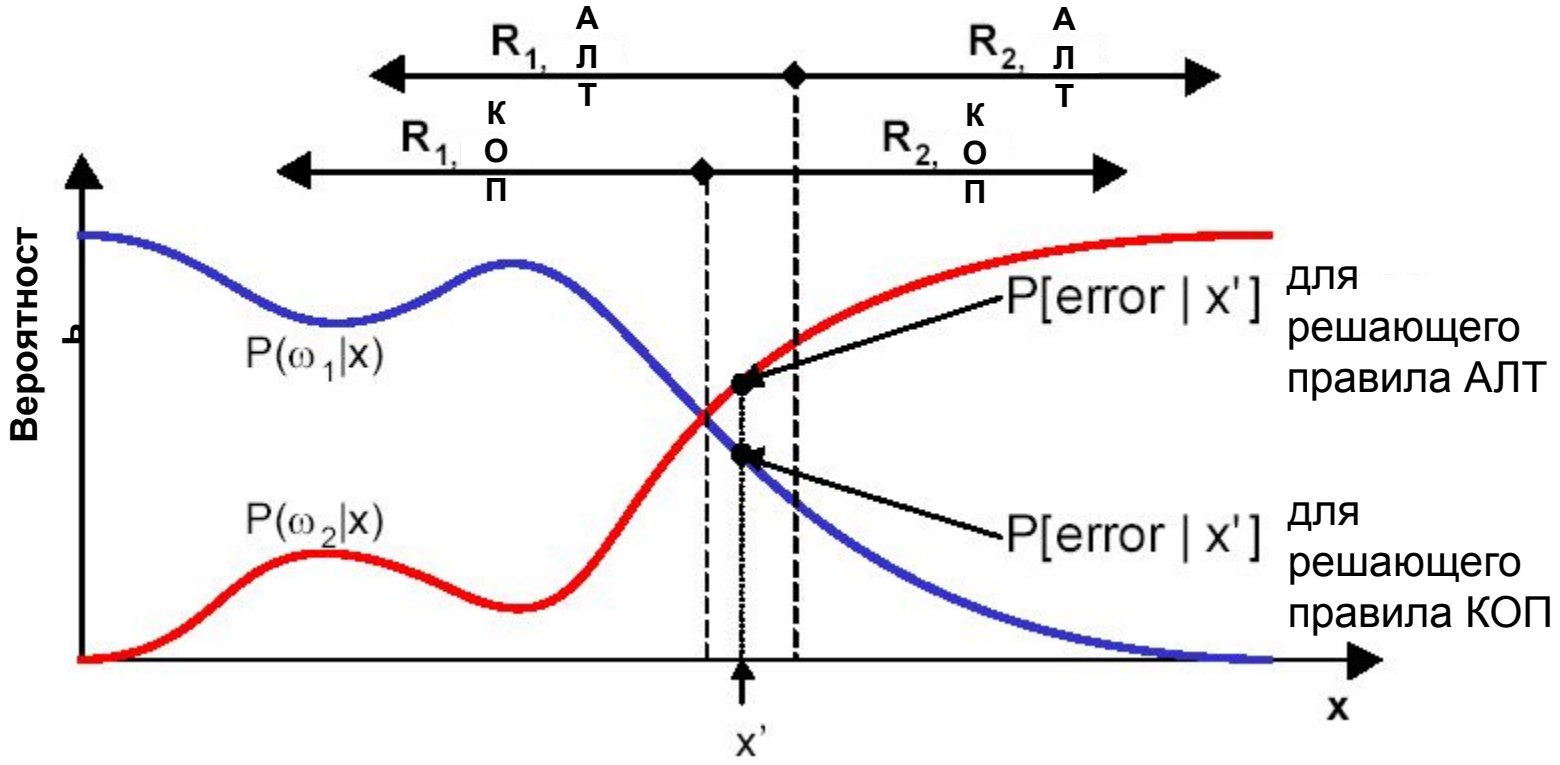
$$P[\text{error}] = \int_{-\infty}^{+\infty} P[\text{error} \mid x] P(x) dx$$

  - Оптимальное решающее правило – правило, минимизирующее P[error | x] для любого x, чтобы минимизировать весь интеграл

  - В каждой точке x’ вероятность P[error | x’] = $P(\omega_i|x')$, если выбран другой класс ω_j

Вероятность ошибки

- В каждой точке x’ вероятность P[error | x’] = $P(\omega_i|x')$, если выбран другой класс ω_j



Вероятность ошибки

Для любой задачи минимальная вероятность ошибки достигается, если в качестве решающего правила используется критерий отношения правдоподобия.

Эта вероятность ошибки называется **байесовским уровнем ошибки** и является **наилучшим** значением среди всех возможных классификаторов.

Байесовский риск

- До сих пор полагалось, что цена ошибочного отнесения к классу ω_1 образца, принадлежащего классу ω_2, совпадает с ценой противоположной ошибки. В общем случае это не верно.
  - Ошибочное отнесение больного раком к здоровым имеет несравненно более печальные последствия, чем обратная ошибка

- Это формализуется введением функции стоимости ошибок C_{ij}
  - C_{ij} – цена отнесения образца к классу ω_i если на самом деле он принадлежит классу ω_j

- Байесовский риск определяется как мат. ожидание стоимости

$$\Re = E[C] = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} \cdot P[\text{ выбран } \omega_i \text{ при } x \in \omega_j] = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} \cdot P[x \in R_i \mid \omega_j] \cdot P[\omega_j]$$

Байесовский риск

- Какое решающее правило минимизирует байесовский риск?
  - Заметим, что

$$P[x \in R_i \mid \omega_j] = \int_{R_i} P(x \mid \omega_j)dx$$

  - Байесовский риск выражается как

$$\Re = \int_{R_1} [C_{11} \cdot P[\omega_1] \cdot P(x \mid \omega_1) + C_{12} \cdot P[\omega_2] \cdot P(x \mid \omega_2)]dx +$$
$$\int_{R_2} [C_{21} \cdot P[\omega_1] \cdot P(x \mid \omega_1) + C_{22} \cdot P[\omega_2] \cdot P(x \mid \omega_2)]dx$$

  - Для любого правдоподобия

$$\int_{R_1} P(x \mid \omega_i)dx + \int_{R_2} P(x \mid \omega_i)dx = \int_{R_1 \cup R_2} P(x \mid \omega_i)dx = 1$$

Байесовский риск

- Подставим последнее уравнение в выражение байесовского риска

$$\begin{aligned}\Re = {} & C_{11}P[\omega_1]\int_{R_1}P(x\mid\omega_1)dx + C_{12}P[\omega_2]\int_{R_1}P(x\mid\omega_2)dx + \\ & + C_{21}P[\omega_1]\int_{R_2}P(x\mid\omega_1)dx + C_{22}P[\omega_2]\int_{R_2}P(x\mid\omega_2)dx + \\ & + C_{21}P[\omega_1]\int_{R_1}P(x\mid\omega_1)dx + C_{22}P[\omega_2]\int_{R_1}P(x\mid\omega_2)dx + \\ & - C_{21}P[\omega_1]\int_{R_1}P(x\mid\omega_1)dx - C_{22}P[\omega_2]\int_{R_1}P(x\mid\omega_2)dx\end{aligned}$$

- Избавимся от интегралов по R_2

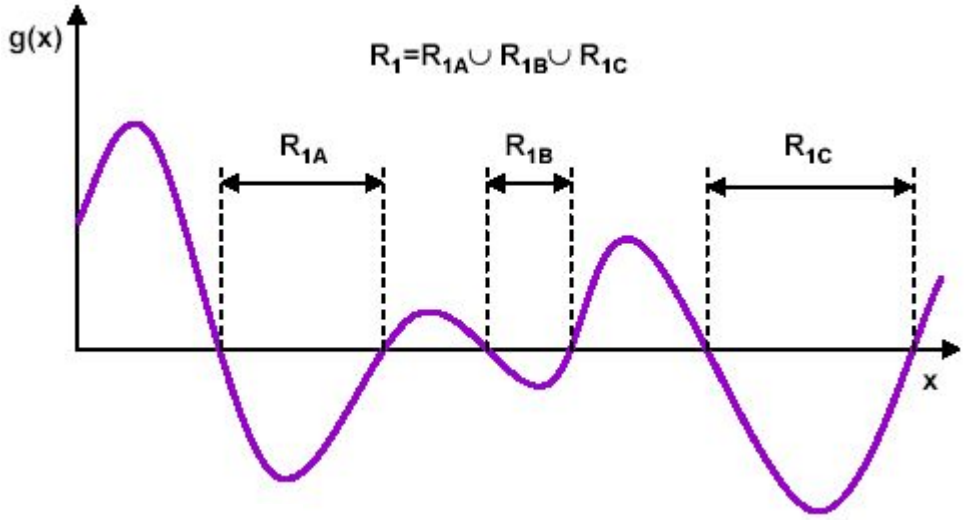
$$\begin{aligned}\Re = {} & C_{21}P[\omega_1] + C_{22}P[\omega_2] + \\ & + \underbrace{(C_{12}-C_{22})}_{0}P[\omega_2]\int_{R_1}P(x\mid\omega_2)dx - \underbrace{(C_{21}-C_{11})}_{0}P[\omega_1]\int_{R_1}P(x\mid\omega_1)dx\end{aligned}$$

- Т.к. первые два слагаемых – константы, ищем R_1, минимизирующее остальное:

$$\begin{aligned}R_1 &= \arg\min\left\{\int_{R_1}\left[(C_{12}-C_{22})P[\omega_2]P(x\mid\omega_2)-(C_{21}-C_{11})P[\omega_1]P(x\mid\omega_1)\right]dx\right\} \\ &= \arg\min\left\{\int_{R_1}g(x)dx\right\}\end{aligned}$$

Байесовский риск

- На уровне интуиции: к каким областям R_1 приводит минимизация байесовского риска?
  - Отыскивается R_1, минимизирующая интеграл $\int_{R_1} g(x)dx$
  - Иными словами, ищутся области, в которых $g(x) < 0$



  - Таким образом, R_1 выбирается так, чтобы

$$(C_{21} - C_{11})P[\omega_1]P(x \mid \omega_1) \overset{\omega_1}{>} (C_{12} - C_{22})P[\omega_2]P(x \mid \omega_2)$$

Байесовский риск

- На уровне интуиции: к каким областям R_1 приводит минимизация байесовского риска?
  - Отыскивается R_1, минимизирующая интеграл $\int_{R_1} g(x)dx$
  - Иными словами, ищутся области, в которых $g(x) < 0$
  - Таким образом, R_1 выбирается так, чтобы

$$(C_{21} - C_{11})P[\omega_1]P(x \mid \omega_1) \overset{\omega_1}{>} (C_{12} - C_{22})P[\omega_2]P(x \mid \omega_2)$$

  - Или, после нехитрых преобразований:

$$\frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{(C_{12} - C_{22})}{(C_{21} - C_{11})} \frac{P[\omega_2]}{P[\omega_1]}$$

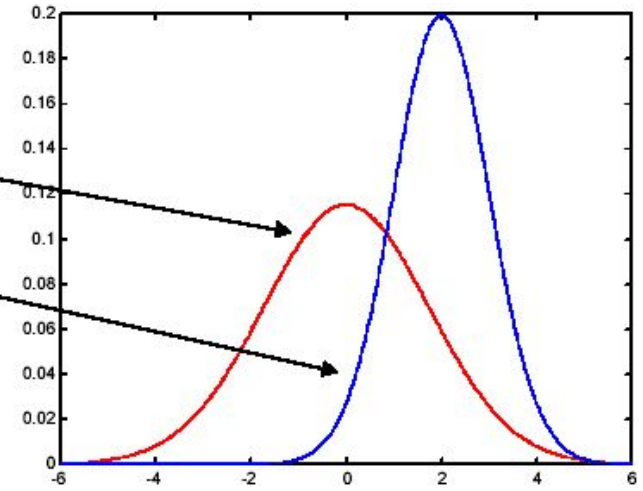
  - Итак, минимизация байесовского риска снова приводит к **критерию отношения правдоподобий**

Байесовский риск: пример

- Классификация на два класса с функциями правдоподобия

$$P(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}$$

$$P(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$$

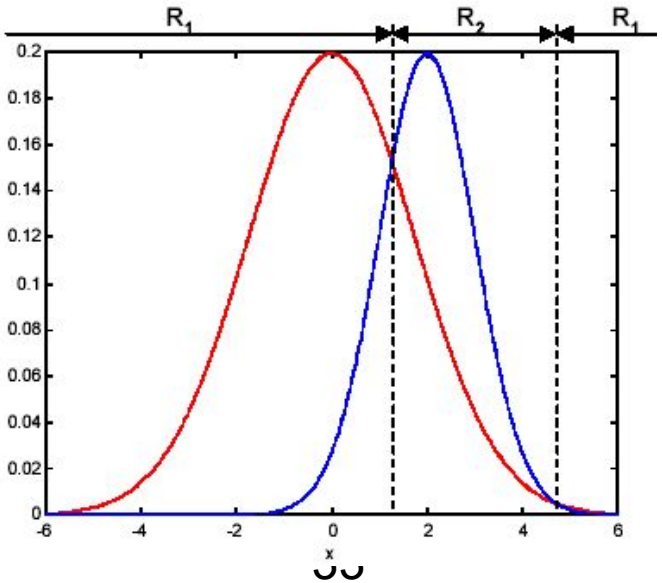


  - Пусть $P(\omega_1) = P(\omega_2) = 0.5$, $C_{11} = C_{22} = 0$, $C_{12} = 1$, $C_{21} = 3^{1/2}$
  - Построить решающее правило, минимизирующее вероятность ошибки

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{1}{\sqrt{3}}$$

$$\frac{e^{-\frac{1}{2}\frac{x^2}{3}}}{e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1 \qquad -\frac{1}{2}\frac{x^2}{3} + \frac{1}{2}(x-2)^2 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$

$$2x^2 - 12x + 12 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \Rightarrow x = 4.73, 1.27$$



Вариации критерия отношения правдоподобия

- Решающее правило КОП, минимизирующее байесовский риск обычно называется **критерием Байеса**

$$\Lambda(x) = \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{(C_{12} - C_{22})}{(C_{21} - C_{11})} \frac{P[\omega_2]}{P[\omega_1]}$$

Критерий Байеса

- В частном случае, при минимизации вероятности ошибки, решающее правило называется **максимальным апостериорным критерием** (МАК)

$$C_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \Rightarrow \Lambda(x) = \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)} \Leftrightarrow \frac{P(\omega_1 \mid x)}{P(\omega_2 \mid x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

Максимальный апостериорный критерий

- В случае равных априорных вероятностей $P(\omega_i) = ½$ решающее правило называется **критерием максимального правдоподобия**

$$\left.\begin{aligned} C_{ij} &= \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \\ P(\omega_i) &= \frac{1}{C} \quad \forall i \end{aligned}\right\} \Rightarrow \Lambda(x) = \frac{P(x \mid \omega_1)}{P(x \mid \omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

Критерий максимального правдоподобия

Правило минимизации P[error] для многоклассовых задач

- Правило минимизации вероятности ошибки P[error] легко обобщается на случай многих классов
  - Для простоты вводится дополнительная вероятность – вероятность правильного выбора класса:

    P[error] = 1 – P[correct]
  - Вероятность корректного распознавания

$$P[\text{correct}] = \sum_{i=1}^{C} P(\omega_i) \int_{R_i} P(x|\omega_i)dx$$

  - Задача минимизации P[error] эквивалентна максимизации P[correct]
  - P[correct] в терминах апостериорных вероятностей:

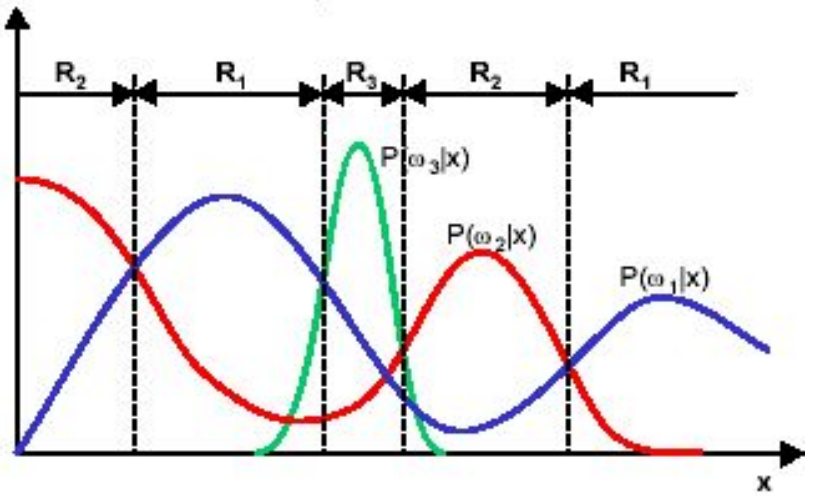
$$P[\text{correct}] = \sum_{i=1}^{C} P(\omega_i) \int_{R_i} P(x|\omega_i)dx = \sum_{i=1}^{C} \int_{R_i} P(x|\omega_i)P(\omega_i)dx = \sum_{i=1}^{C} \underbrace{\int_{R_i} P(\omega_i|x)P(x)dx}_{\Im_i}$$

Правило минимизации P[error] для многоклассовых задач

- Правило минимизации вероятности ошибки P[error] легко обобщается на случай многих классов
  - P[correct] в терминах апостериорных вероятностей:

$$P[\text{correct}] = \sum_{i=1}^{C} P(\omega_i) \int_{R_i} P(x|\omega_i)dx = \sum_{i=1}^{C} \int_{R_i} P(x|\omega_i)P(\omega_i)dx = \sum_{i=1}^{C} \underbrace{\int_{R_i} P(\omega_i|x)P(x)dx}_{J_i}$$

  - Для максимизации P[correct] нужно максимизировать каждый из интегралов J_i. Интегралы максимизируются выбором классов ω_i, дающих максимум $P[\omega_i|x]$, т.е. определением области R_i, в которой $P[\omega_i|x]$ максимальна.



- Таким образом, правило минимизации P[error] – максимальный апостериорный критерий

Минимизация байесовского риска для многоклассовых задач

- Новые обозначения
  - α_i – решение о выборе класса ω_i
  - $\alpha(x)$ – общее правило выбора, отображающее вектор x на классы ω_i: $\alpha(x) \to \{\alpha_1, \alpha_2, \ldots, \alpha_C\}$

$$P[\text{correct}] = \sum_{i=1}^{C} P(\omega_i) \int_{R_i} P(x|\omega_i)dx = \sum_{i=1}^{C} \int_{R_i} P(x|\omega_i)P(\omega_i)dx = \sum_{i=1}^{C} \underbrace{\int_{R_i} P(\omega_i|x)P(x)dx}_{\Im_i}$$

- Условный риск $R(\alpha_i|x)$ отнесения вектора x к классу ω_i есть

$$\Re(\alpha(x) \to \alpha_i) = \Re(\alpha_i \mid x) = \sum_{j=1}^{C} C_{ij} P(\omega_j \mid x)$$

- Байесовский риск, связанный с решающим правилом $\alpha(x)$ есть

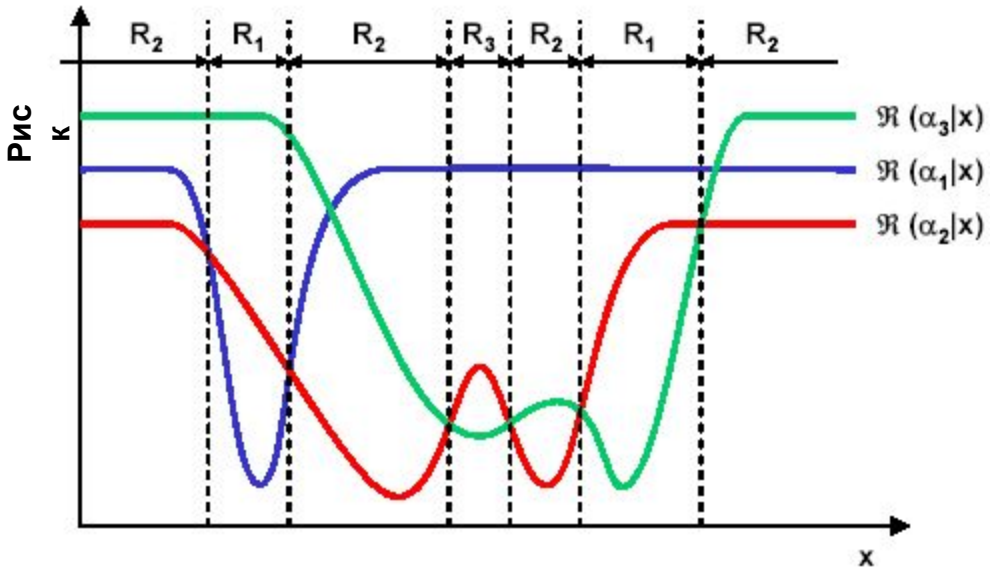
$$\Re(\alpha(x)) = \int \Re(\alpha(x) \mid x) P(x)dx$$

Минимизация байесовского риска для многоклассовых задач

- Байесовский риск, связанный с решающим правилом $\alpha(x)$ есть

$$\Re(\alpha(x)) = \int \Re(\alpha(x) \mid x) P(x) dx$$

- Для минимизации этого выражения необходимо минимизировать условный риск $R(\alpha(x)|x)$ в каждой точке пространства x, что эквивалентно выбору класса ω_i, такого, что $R(\alpha_i|x)$ минимален

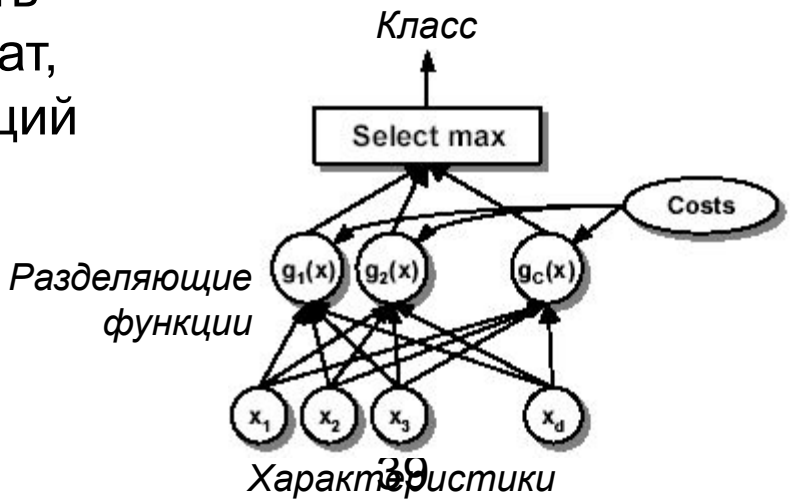


Разделяющие функции

- Все решающие правила, рассмотренные в этой лекции, имеют одинаковую структуру
  - В каждой точке x пространства признаков выбрать класс ω_i такой, что минимизируется (максимизируется) некоторая мера $\mathbf{g_i(x)}$
- Эта структура может быть формализована с помощью множества разделяющих функций $g_i(x)$, i=1,…,C и следующего решающего правила

> “отнести x к классу ω_i, если $g_i(x)>g_j(x) \quad \forall j \neq i$”

- Тогда решающее правило может быть визуализировано как сеть или автомат, вычисляющий C разделяющих функций и выбирающий класс, соответствующий наибольшему значению



Разделяющие функции

- Основные критерии как разделяющие функции

Критерий	Разделяющая функция
Байеса	$g_i(x)=-\Re(\alpha_i \mid x)$
Максимильный апостериорный	$g_i(x)=P(\omega_i \mid x)$
Максимального правдоподобия	$g_i(x)=P(x \mid \omega_i)$

Байесовские классификаторы для нормально распределенных классов

- Для случая нормально распределенных классов критерий минимизации ошибки (максимизации апостериорной вероятности, МАВ)

> “отнести x к классу ω_i, если $g_i(x) > g_j(x) \quad \forall j \neq i$”

где $g_i(x) = P(\omega_i \mid x)$, может быть существенно упрощен

Байесовские классификаторы для нормально распределенных классов

- Выражения для гауссовых плотностей в общем виде
  - Плотность многомерного нормального распределения

$$f_x(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right]$$

  - Преобразованная разделяющая функция критерия MAB (по формуле Байеса)

$$g_i(x) = P(\omega_i \mid x) = \frac{P(x \mid \omega_i)P(\omega_i)}{P(x)} = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right] P(\omega_i)\frac{1}{P(x)}$$

  - Исключая постоянные члены, получим

$$g_i(x) = |\Sigma_i|^{-1/2} \exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right] P(\omega_i)$$

  - После логарифмирования

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i))$$

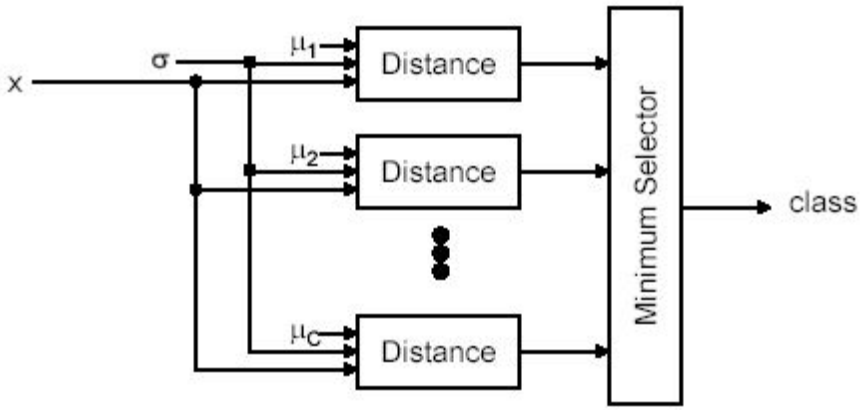
    - Это выражение называется **квадратичной разделяющей функцией**

Случай 1: $\Sigma_i = \sigma^2 I$

- Случай возникает. когда характеристики образцов статистически независимы и имеют одинаковую вариацию по всем классам

$$g_i(x) = -\frac{1}{2\sigma^2}\left(-2\mu_i^T x + \mu_i^T \mu_i\right) + \log(P(\omega_i)) = w_i^T x + w_{i0}$$

$$\text{where} \begin{cases} w_i = \dfrac{\mu_i}{\sigma^2} \\ w_{i0} = -\dfrac{1}{2\sigma^2}\mu_i^T \mu_i + \log(P(\omega_i)) \end{cases}$$



  - Т.к. $g_i(x)$ линейны, границы классов – гиперплоскости
  - Если априорные вероятности равны, то

$$g_i(x) = -\frac{1}{2\sigma^2}(x - \mu_i)^T (x - \mu_i)$$

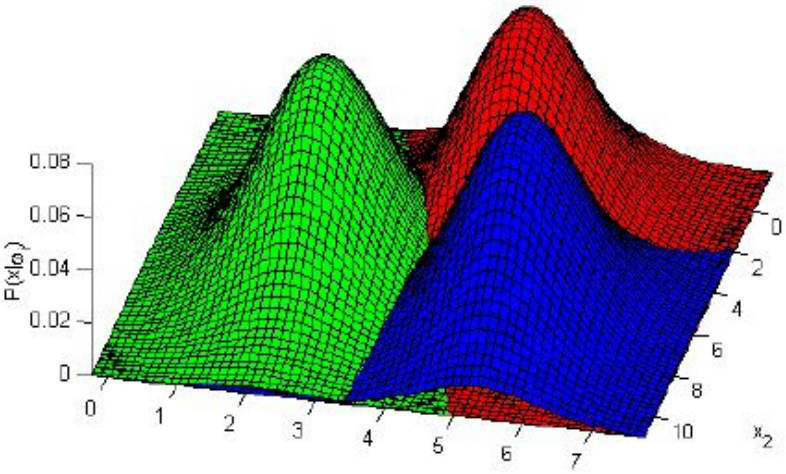
- Это – классификатор по минимальному расстоянию (по ближайшему среднему)
- ГМТ с постоянной вероятностью – гиперсферы
- Для единичной дисперсии, расстояние становится еклидовым

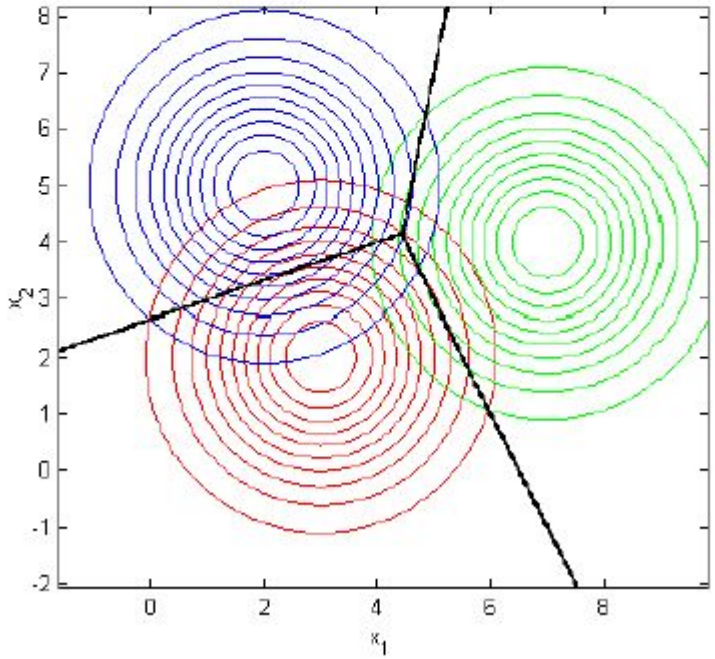
Случай 1: $\Sigma_i = \sigma^2 I$, пример

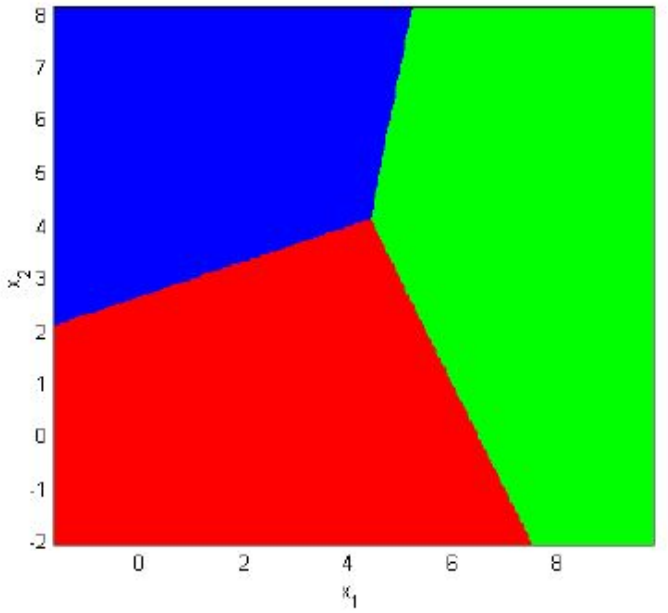
- Двумерная задача, три класса со следующими средними и ковариациями:

$$\mu_1 = [3 \quad 2]^T \quad \mu_2 = [7 \quad 4]^T \quad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$







Случай 2: $\Sigma_i = \Sigma$ (Σ – диагональная)

- Классы по-прежнему имеют одинаковую матрицу ковариации, но характеристики имеют разные дисперсии

$$g_i(x) = -\frac{1}{2}\sum_{k=1}^{N}\frac{2x[k]\mu_i[k] + \mu_i[k]^2}{\sigma_k^2} - \frac{1}{2}\log\prod_{k=1}^{N}\sigma_k^2 + \log(P(\omega_i))$$

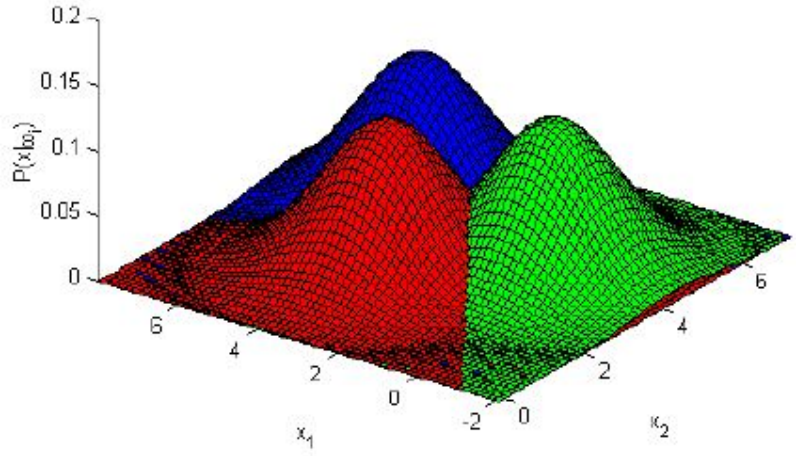
- Функция линейна, границы классов – гиперплоскости
- ГМТ точек с одинаковой вероятностью – гиперэллипсоиды
- Единственное отличие от предыдущего случая – нормализация осей

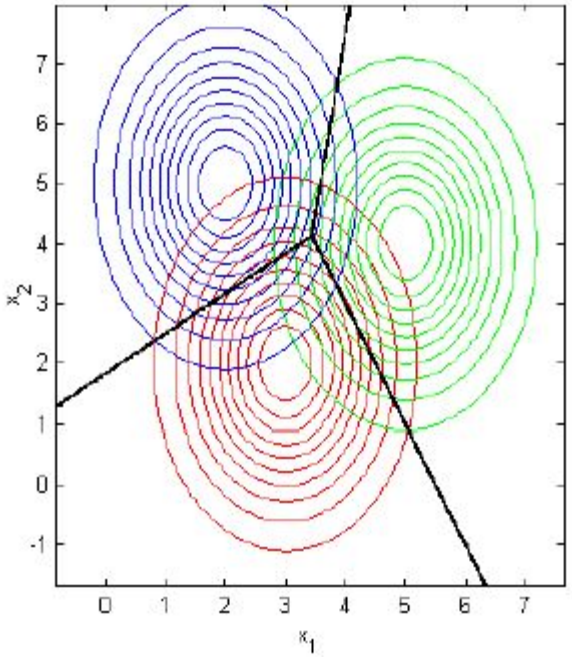
Случай 2: $\Sigma_i = \Sigma$ (Σ – диагональная), пример

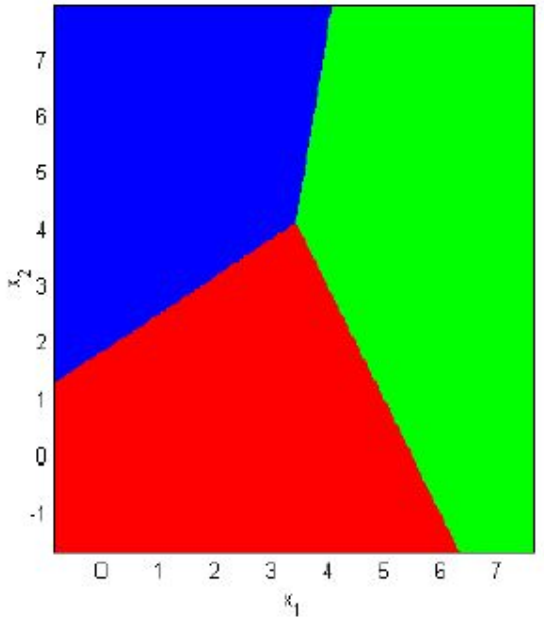
- Двумерная задача, три класса со следующими средними и ковариациями:

$$\mu_1 = [3 \quad 2]^T \quad \mu_2 = [5 \quad 4]^T \quad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$







Случай 3: $\Sigma_i = \Sigma$ (Σ – не диагональная)

- Классы по-прежнему имеют одинаковую матрицу ковариации, но матрица – не диагональная

$$g_i(x) = w_i^T x + w_{i0}$$

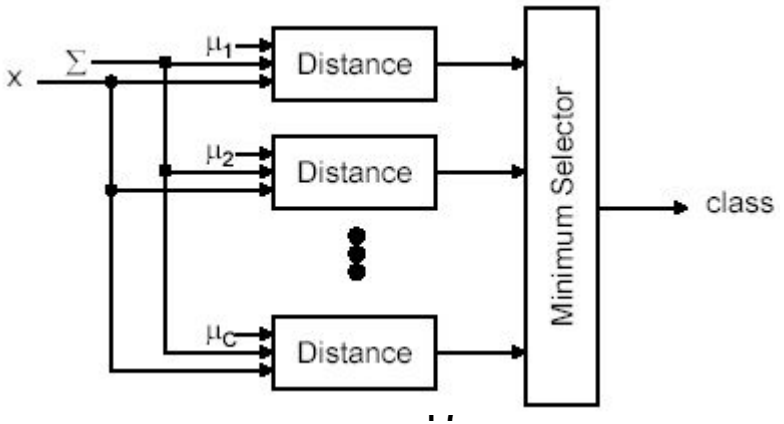
$$\text{where} \begin{cases} w_i = \Sigma^{-1}\mu_i \\ w_{i0} = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log P(\omega_i) \end{cases}$$

- Функция линейна, границы классов – гиперплоскости
- ГМТ точек с одинаковой вероятностью – гиперэллипсоиды, натянутые на собственные вектора матрицы ковариации
- При равных априорных вероятностях – классифкатор по минимальному расстоянию Махаланобиса:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)$$

Расстояние Махаланобиса

$$\|x-y\|^2_{\Sigma^{-1}} = (x-y)^T \Sigma^{-1}(x-y)$$

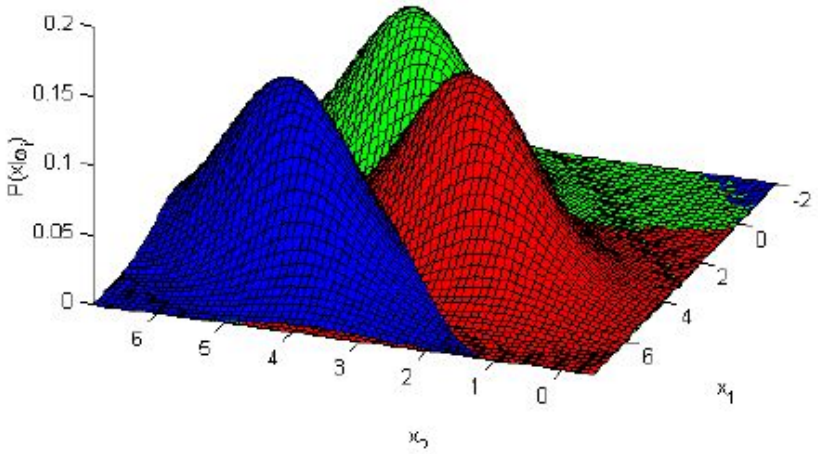


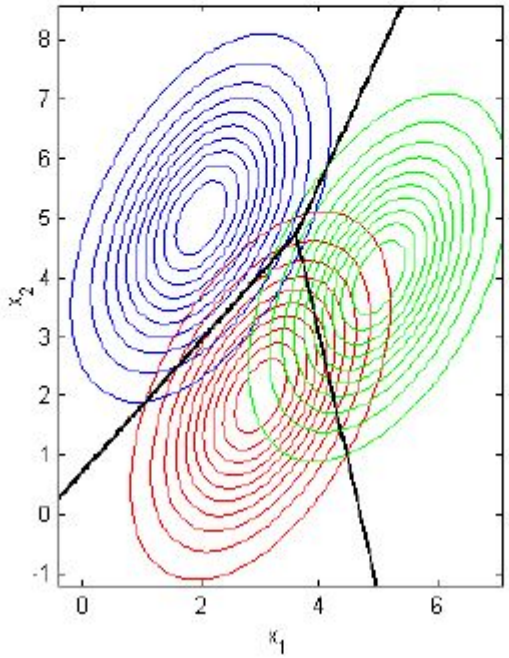
Случай 3: $\Sigma_i = \Sigma$ (Σ – не диагональная), пример

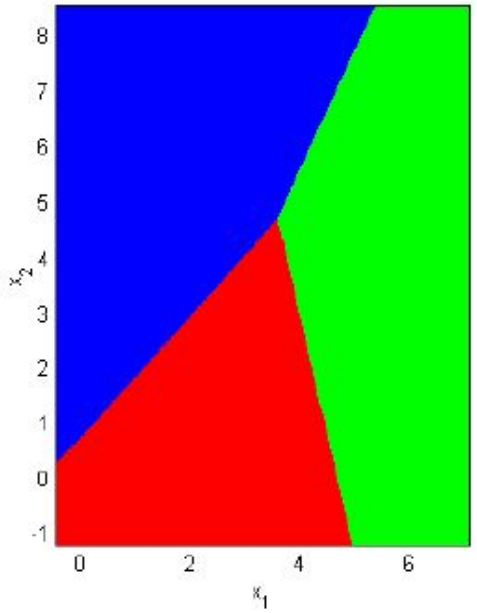
- Двумерная задача, три класса со следующими средними и ковариациями:

$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 2 \end{bmatrix}$$







Случай 4: $\Sigma_i = \sigma_i^2 I$

- Классы имеют разные матрицы ковариации, которые пропорциональны единичной матрице

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log\left(\left|\Sigma_i\right|\right) + \log(P(\omega_i)) =$$
$$= -\frac{1}{2}(x-\mu_i)^T \sigma_i^{-2}(x-\mu_i) - \frac{1}{2}N\log\left(\sigma_i^2\right) + \log(P(\omega_i))$$

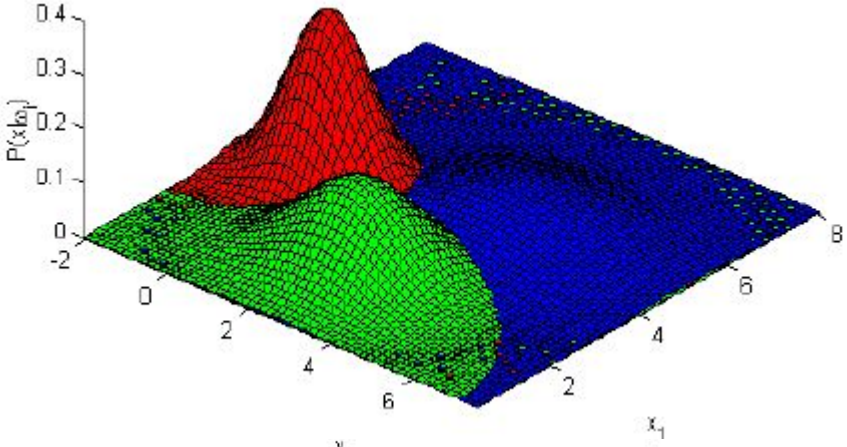
- Границы классов – гиперэллипсоиды
- ГМТ точек с одинаковой вероятностью – гиперсферы, определяемые осями пространства характеристик

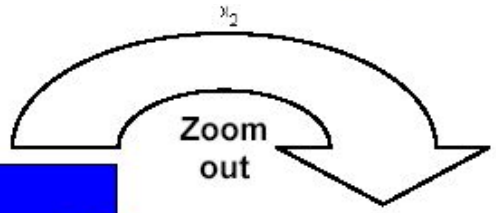
Случай 4: $\Sigma_i = \sigma_i^2 I$, пример

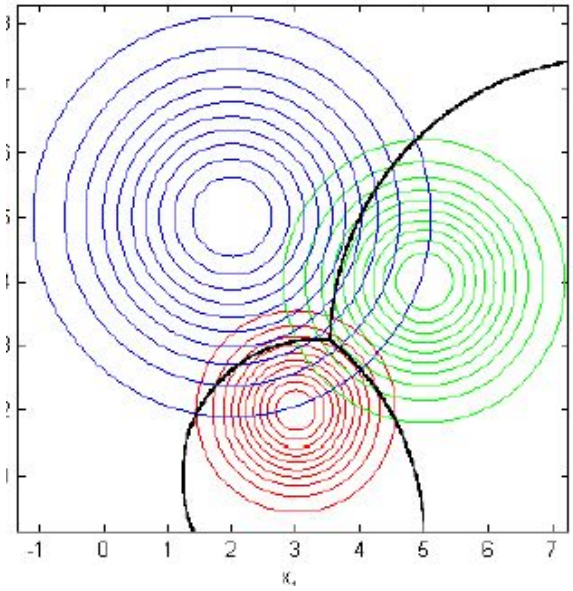
- Двумерная задача, три класса со следующими средними и ковариациями:

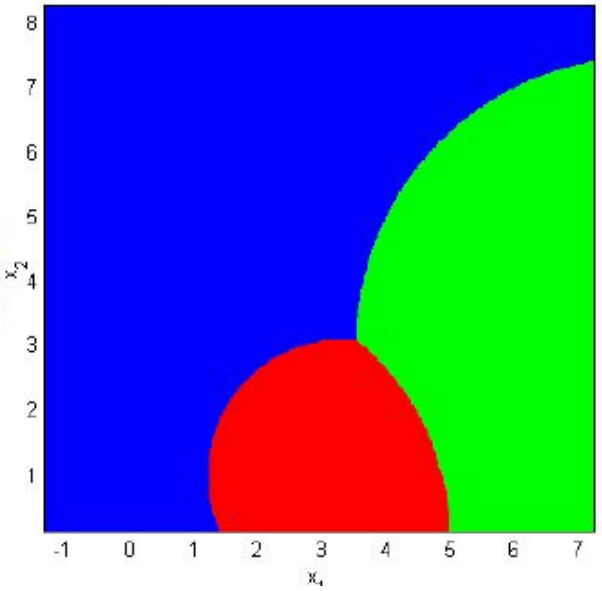
$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

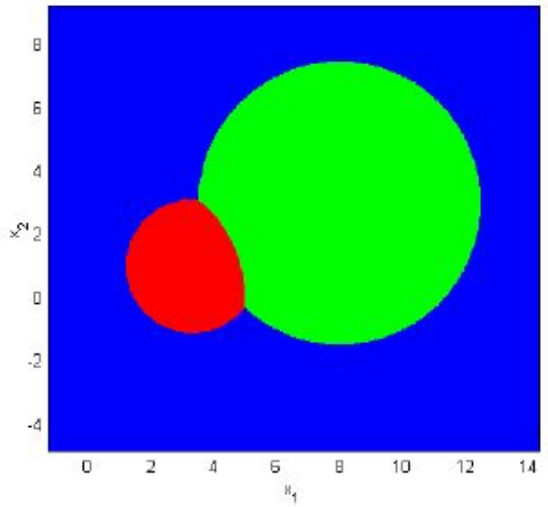
$$\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$











Случай 5: $\Sigma_i \neq \Sigma_j$ (общий случай)

- Для общего случая разделяющая функция выглядит так:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i))$$

- Или, после преобразований:

$$g_i(x) = x^T W_i x + w_i^T x + w_{i0}$$

$$\text{where} \begin{cases} W_i = -\frac{1}{2}\Sigma_i^{-1} \\ w_i = \Sigma_i^{-1}\mu_i \\ w_{i0} = -\frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i)) \end{cases}$$

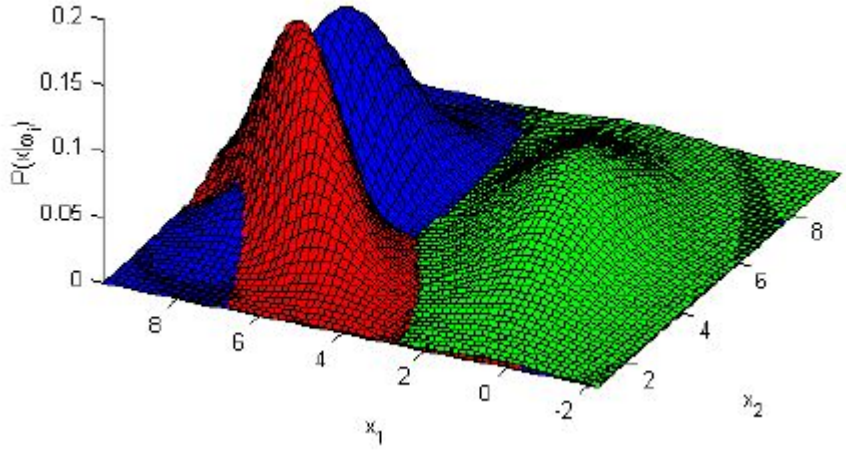
- Границы классов – гиперэллипсоиды и гиперпараболоиды
- ГМТ точек с одинаковой вероятностью – гиперэллипсоиды, определяемые собственными векторами матриц Σ_i для каждого из классов в отдельности

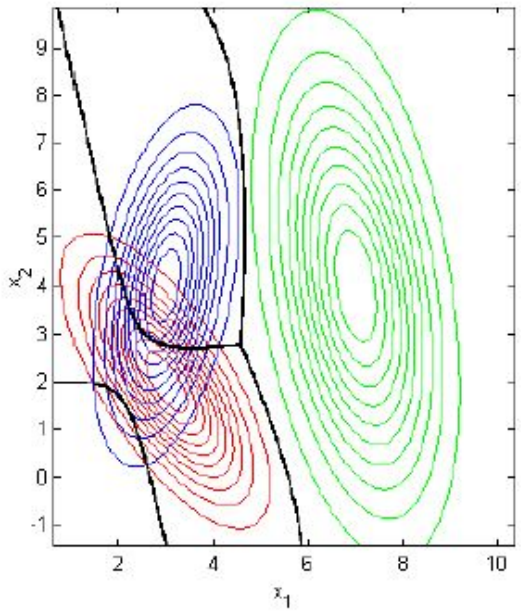
Случай 5: $\Sigma_i \neq \Sigma_j$ (общий случай), пример

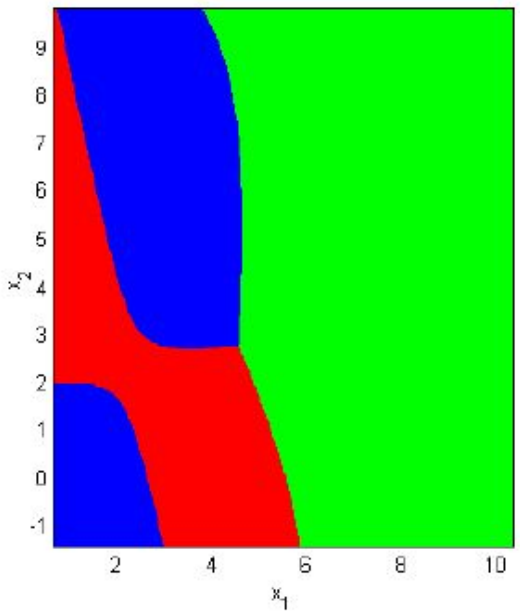
- Двумерная задача, три класса со следующими средними и ковариациями:

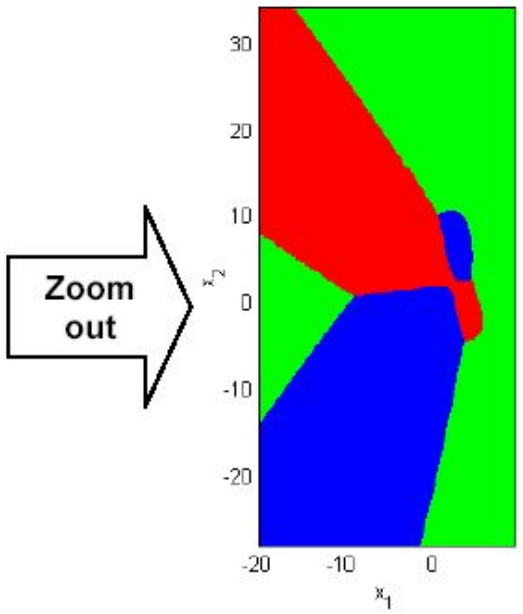
$$\mu_1 = [3 \quad 2]^T \qquad \mu_2 = [5 \quad 4]^T \qquad \mu_3 = [2 \quad 5]^T$$

$$\Sigma_1 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 1 & -1 \\ -1 & 7 \end{bmatrix} \qquad \Sigma_3 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 3 \end{bmatrix}$$









Численный пример

- Построить линейную разделяющую функцию для трехмерной двухклассовой задачи распознавания по следующим данным:

$$\mu_1 = [0 \quad 0 \quad 0]^T;\ \mu_2 = [1 \quad 1 \quad 1]^T;\ \Sigma_1 = \Sigma_2 = \begin{bmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{bmatrix};\ p(\omega_2) = 2p(\omega_1)$$

- Решение

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + \log P(\omega_i) = -\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \\ z-\mu_z \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \\ z--\mu_z \end{bmatrix} + \log P(\omega_i)$$

$$g_1(x) = -\frac{1}{2}\begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix} + \log\frac{1}{3};\qquad g_2(x) = -\frac{1}{2}\begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix} + \log\frac{2}{3}$$

$$g_1(x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} g_2(x) \Rightarrow -2\left(x^2+y^2+z^2\right) + \log\frac{1}{3} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} -2\left((x-1)^2+(y-1)^2+(z-1)^2\right) + \log\frac{2}{3}$$

$$x+y+z \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \frac{6-\log 2}{4} = 1.32$$

- Классифицировать тестовый образец $x_t = [0.1\ 0.7\ 0.8]^T$

$$0.1+0.7+0.8 = 1.6 \underset{\omega_1}{\overset{\omega_2}{<}} 1.32 \Rightarrow x_u \in \omega_2$$

Некоторые заключения

- Байесовские классификаторы для нормально распределенных классов – квадратичные функции
- Байесовские классификаторы для нормально распределенных классов с одинаковыми ковариационными матрицами – линейные функции
- Классификатор по минимальному расстоянию Махалонобиса оптимален по Байесу, если
  - классы нормально распределены **и**
  - матрицы ковариаций равны **и**
  - априорные вероятности равны
- Классификатор по минимальному евклидову расстоянию оптимален по Байесу, если
  - классы нормально распределены **и**
  - матрицы ковариаций равны и пропорциональны единичной матрице **и**
  - априорные вероятности равны