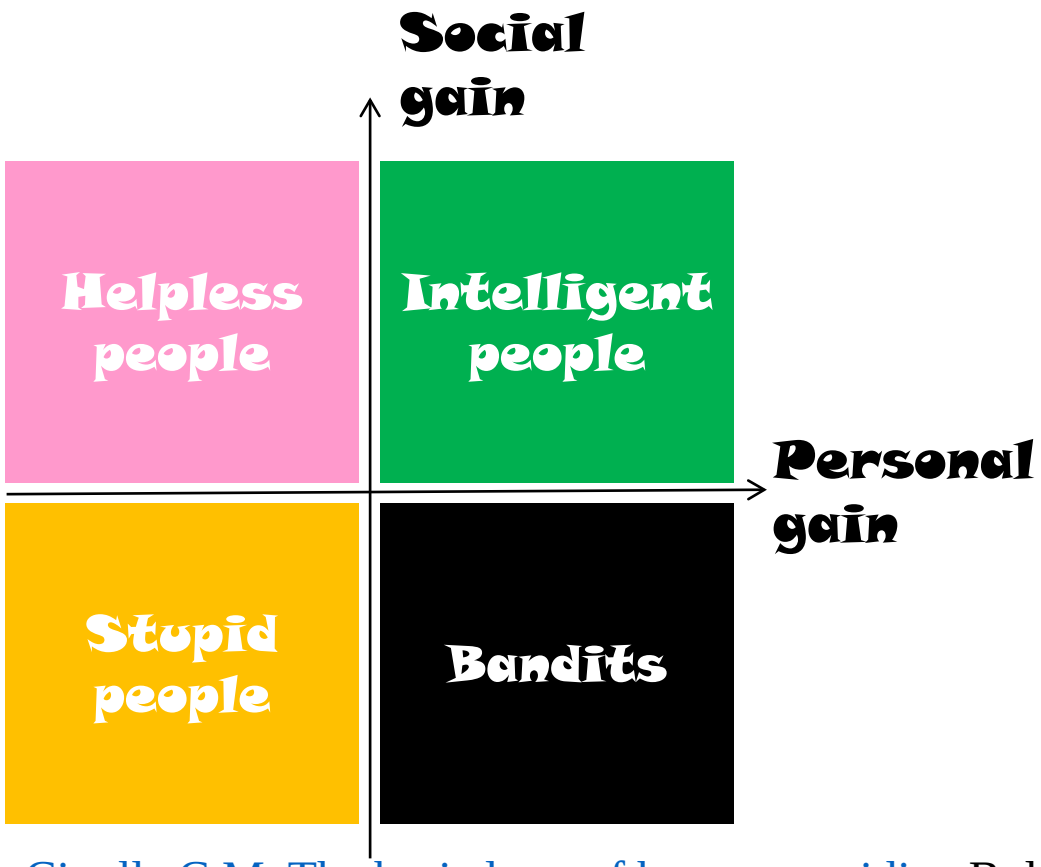
Задача классификации данных



Классификация – нить Ариадны в лабиринте природы.

Жорж Санд

Cipolla C.M. The basic laws of human stupidity. Bologna: il Muli, 2011 (текст)

Содержание

- Основные понятия
- Деревья решений
- **Байесовская классификация**
- Классификация по ближайшим соседям
- Оценка качества классификации
- Ансамблевая классификация

(Наивная) Б*а*йесовская классификация

- Идея
  - Предсказать вероятность принадлежности объекта классу
- Базис
  - Теорема Байеса
- Точность
  - Сопоставима с деревьями решений
- Инкрементальность
  - Возможность пополнения обучающей выборки
- Baseline (линия отсчета)
  - Минимум для сравнения других методов

Теорема Б*а*йеса

Томас Б*а*йес
1702-1761

- $P(A|B)$ – условная вероятность (вероятность наступления события A при условии, что событие B произошло):

  $P(A|B) = \frac{P(AB)}{P(B)}$, $P(B|A) = \frac{P(AB)}{P(A)}$, AB – совместное событие

- $AB \equiv BA$:

$$P(AB) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

- Формула Байеса

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- Формула полной вероятности для $P(B) \neq 0$:

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)$$

$A_1, \dots, A_n$ – несовместные события, $\sum_{i=1}^{n} P(A_i) = 1$

Теорема Байеса: пример

- A – в бензобаке нет бензина
  B – авто не заводится
- $P(B|A) = 1$
- $P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} = \frac{P(A)}{P(B)}$
- Если $P(A) = 0.01, P(B) = 0.02$
  то для случайно выбранной машины
  вероятность того, что в бензобаке нет топлива,
  равна 0.5

Теорема Байеса: пример

- Играют Трактор и АкБарс на поле АкБарс. Кто победит, если
  - Трактор выиграл 65% игр, АкБарс – 35% игр
  - Трактор выиграл 30% игр на чужом поле, АкБарс выиграл 75% игр на своем поле
- В терминах вероятностей
  - $P(\text{Победа} = \text{Трактор}) = 0.65$,
  - $P(\text{Победа} = \text{АкБарс}) = 0.35$
  - $P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{АкБарс}) = 0.75$
  - $P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{Трактор}) = 0.30$
  - Найти $P(\text{Победа} = \text{АкБарс}|\text{Поле} = \text{АкБарс})$

Теорема Байеса: пример

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- В терминах вероятностей
  - $P(\text{Победа} = \text{Трактор}) = 0.65$, $P(\text{Победа} = \text{АкБарс}) = 0.35$
  - $P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{АкБарс}) = 0.75$
  - $P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{Трактор}) = 0.30$
  - Найти $P(\text{Победа} = \text{АкБарс}|\text{Поле} = \text{АкБарс})$
- Применение теоремы Байеса

$$P(\text{Победа} = \text{АкБарс}|\text{Поле} = \text{АкБарс}) = \frac{P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{АкБарс}) \cdot P(\text{Победа} = \text{АкБарс})}{P(\text{Поле} = \text{АкБарс})}$$

$$= \frac{P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{АкБарс}) \cdot P(\text{Победа} = \text{АкБарс})}{P(\text{Поле} = \text{АкБарс}, \text{Победа} = \text{АкБарс}) + P(\text{Поле} = \text{АкБарс}, \text{Победа} = \text{Трактор})} =$$

$$= \frac{P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{АкБарс}) \cdot P(\text{Победа} = \text{АкБарс})}{P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{АкБарс}) \cdot P(\text{Победа} = \text{АкБарс}) + P(\text{Поле} = \text{АкБарс}|\text{Победа} = \text{Трактор}) \cdot \text{P}(\text{Победа} = \text{Трактор})} =$$

$$= \frac{0.75 \cdot 0.35}{0.75 \cdot 0.35 + 0.3 \cdot 0.65} = 0.5738$$

- $\boldsymbol{P}(\text{Победа} = \text{АкБарс}|\text{Поле} = \text{АкБарс}) = \mathbf{57.4\%}$

Применение теоремы Байеса для классификации

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

$\boldsymbol{X}$	Классифицируемый объект	Пациент
$\boldsymbol{A_1, \dots, A_n}$	Атрибуты объекта	(Возр, Пол, t^0)
$\boldsymbol{x_1, \dots, x_n}$	Значения атрибутов объекта	(пожилой, м, высокая)
$\boldsymbol{C_i}$	Классы объектов	Здоров, Болен
$\boldsymbol{H}$	Гипотеза «$X \in C_i$»	Пациент здоров
$P(\boldsymbol{H}\|\boldsymbol{X})$	Апостериорная вероятность H при условии X	Вероятность, что пациент здоров, если это пожилой мужчина с высокой t^0
$P(\boldsymbol{H})$	Априорная вероятность H	Вероятность, что пациент здоров
$P(\boldsymbol{X}\|\boldsymbol{H})$	Апостериорная вероятность X при условии H	Вероятность, что здоровый пациент – это пожилой мужчина с высокой t^0
$P(\boldsymbol{X})$	Априорная вероятность X	Вероятность, что пациент – это пожилой мужчина с высокой t^0

Байесовская классификация

- Принцип классификации

$$X \in C_k \Leftrightarrow C_k = \arg\max_{C_i} P(C_i|X)$$

- Применим теорему Байеса

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

$P(X)$ одинаково для всех классов

- В итоге

$$X \in C_k \Leftrightarrow C_k = \arg\max_{C_i}(P(X|C_i) \cdot P(C_i))$$

Байесовская классификация

- $X \in C_k \Leftrightarrow C_k = \arg\max_{C_i}(P(X|C_i) \cdot P(C_i))$
- $P(C_i)$ – доля кортежей класса C_i в обучающей выборке D
- Для вычисления $P(X|C_i)$ предположим, что атрибуты $A_1, \dots, A_n$ кортежа $X = (x_1, \dots, x_n)$ **независимы** друг от друга. Тогда

$$P(X|C_i) = P(\forall j\ A_j = x_j|C_i) = \prod_{j=1}^{n} P(A_j = x_j|C_i)$$

Байесовская классификация

- $P(X|C_i) = \prod_{j=1}^{n} P(A_j = x_j|C_i) = \prod_{j=1}^{n} \frac{|\{X \in C_i | X.A_j = x_j\}|}{|C_i|}$
- В итоге:

$$X \in C_k \Leftrightarrow C_k = \arg\max_{C_i}(P(C_i) \cdot P(X|C_i))$$

$$X \in C_k \Leftrightarrow C_k = \arg\max_{C_i}\left(\frac{|C_i|}{|D|} \cdot \prod_{j=1}^{n} \frac{|\{X \in C_i | X.A_j = x_j\}|}{|C_i|}\right)$$

Доля объектов класса C_i в обучающей выборке

Доля объектов класса C_i, имеющих в атрибуте A_j значение x_j

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая) **Класс=?**

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(C_i)$
  - $P(\text{Здоров}) = 9/14 = 0.643$
  - $P(\text{Болен}) = 5/14 = 0.357$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(*Возраст*=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 2/9 = 0.222$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(*Возраст*=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Молодой}|\text{Болен}) = 3/5 = 0.6$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, *Пол*=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Мужчина}|\text{Здоров}) = 6/9 = 0.667$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, *Пол*=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Мужчина}|\text{Здоров}) = 0.667$
  - $P(\text{Мужчина}|\text{Болен}) = 1/5 = 0.2$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, Пол=М, *Гемоглобин*=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Мужчина}|\text{Здоров}) = 0.667$
  - $P(\text{Мужчина}|\text{Болен}) = 0.2$
  - $P(\text{Норма}|\text{Здоров}) = 4/9 = 0.444$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, Пол=М, *Гемоглобин*=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Норма}|\text{Здоров}) = 0.444$
  - $P(\text{Мужчина}|\text{Здоров}) = 0.667$
  - $P(\text{Мужчина}|\text{Болен}) = 0.2$
  - $P(\text{Норма}|\text{Болен}) = 2/5 = 0.4$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Мужчина}|\text{Здоров}) = 0.667$
  - $P(\text{Мужчина}|\text{Болен}) = 0.2$
  - $P(\text{Норма}|\text{Здоров}) = 0.444$
  - $P(\text{Норма}|\text{Болен}) = 0.4$
  - $P(\text{Высокая}|\text{Здоров}) = 6/9 = 0.667$

Пример

Возраст	*Пол*	*Гемо-глобин*	t^0	***Класс***
Молодой	Ж	Высокий	Высокая	Болен
Молодой	Ж	Высокий	Норма	Болен
Средний	Ж	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Высокая	Здоров
Пожилой	М	Низкий	Высокая	Здоров
Пожилой	М	Низкий	Норма	Болен
Средний	М	Низкий	Норма	Здоров
Молодой	Ж	Норма	Высокая	Болен
Молодой	М	Низкий	Высокая	Здоров
Пожилой	М	Норма	Высокая	Здоров
Молодой	М	Норма	Норма	Здоров
Средний	Ж	Норма	Норма	Здоров
Средний	М	Высокий	Высокая	Здоров
Пожилой	Ж	Норма	Норма	Болен

- X=(Возраст=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Мужчина}|\text{Здоров}) = 0.667$
  - $P(\text{Мужчина}|\text{Болен}) = 0.2$
  - $P(\text{Норма}|\text{Здоров}) = 0.444$
  - $P(\text{Норма}|\text{Болен}) = 0.4$
  - $P(\text{Высокая}|\text{Здоров}) = 0.667$
  - $P(\text{Высокая}|\text{Болен}) = 2/5 = 0.4$

Пример

- X=(Возраст=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(C_i)$
  - $P(\text{Здоров}) = 0.643$
  - $P(\text{Болен}) = 0.357$
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Мужчина}|\text{Здоров}) = 0.667$
  - $P(\text{Норма}|\text{Здоров}) = 0.444$
  - $P(\text{Высокая}|\text{Здоров}) = 0.667$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Мужчина}|\text{Болен}) = 0.2$
  - $P(\text{Норма}|\text{Болен}) = 0.4$
  - $P(\text{Высокая}|\text{Болен}) = 0.4$

- $X \in C_k \Leftrightarrow$
$$C_k = \arg\max_{C_i}(P(X|C_i) \cdot P(C_i))$$
- **Класс=Здоров**
$P(X|\text{Здоров}) \cdot P(\text{Здоров})$
$= (0.222 \cdot 0.667 \cdot 0.444 \cdot 0.667) \cdot 0.643$
- **Класс=Болен**
$P(X|\text{Болен}) \cdot P(\text{Болен}) = (0.6 \cdot 0.2 \cdot 0.4 \cdot 0.4) \cdot 0.357$

Пример

- X=(Возраст=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=?**)
- $P(C_i)$
  - $P(\text{Здоров}) = 0.643$
  - $P(\text{Болен}) = 0.357$
- $P(X|C_i)$
  - $P(\text{Молодой}|\text{Здоров}) = 0.222$
  - $P(\text{Мужчина}|\text{Здоров}) = 0.667$
  - $P(\text{Норма}|\text{Здоров}) = 0.444$
  - $P(\text{Высокая}|\text{Здоров}) = 0.667$
  - $P(\text{Молодой}|\text{Болен}) = 0.6$
  - $P(\text{Мужчина}|\text{Болен}) = 0.2$
  - $P(\text{Норма}|\text{Болен}) = 0.4$
  - $P(\text{Высокая}|\text{Болен}) = 0.4$

- $X \in C_k \Leftrightarrow$
  $$C_k = \arg\max_{C_i}(P(X|C_i) \cdot P(C_i))$$
- **Класс=Здоров**
  $P(X|\text{Здоров}) \cdot P(\text{Здоров})$
  $= (0.222 \cdot 0.667 \cdot 0.444 \cdot 0.667) \cdot 0.643$
  $= \mathbf{0.028}$
- **Класс=Болен**
  $P(X|\text{Болен}) \cdot P(\text{Болен})$
  $= (0.6 \cdot 0.2 \cdot 0.4 \cdot 0.4) \cdot 0.357 = \mathbf{0.007}$
- **Итог**
  X=(Возраст=Молодой, Пол=М, Гемоглобин=Норма, t^0=Высокая, **Класс=Здоров**)

Байесовская классификация: не все хорошо

- Нулевая вероятность
  - Как быть, если в обучающей выборке нет объектов с таким же значением атрибута A_j, как у классифицируемого объекта (тогда $P(X|C_i)=0$)?
- Атрибуты с непрерывными значениями
  - Как вычислять апостериорную вероятность $P(X|C_i)$, если атрибут A_j имеет непрерывные (а не дискретные) значения?
- Независимость атрибутов
  - Насколько важно предположение о независимости атрибутов $A_1, \ldots, A_n$?

Проблема нулевой вероятности

- $X \in C_k \Leftrightarrow C_k = \arg\max_{C_i} P(C_i) \cdot P(X|C_i) = \arg\max_{C_i} P(C_i) \cdot \prod_{j=1}^{n} \frac{|\{\boldsymbol{X \in C_i | X.A_j = x_j}\}|}{|C_i|}$
- В обучающей выборке нет объектов с таким же значением атрибута A_j, как у классифицируемого объекта
  - $\exists x_j \left\{\boldsymbol{X \in C_i | X.A_j = x_j}\right\} = \emptyset$
  - $P(X|C_i) = 0 \implies X \notin C_i$, хотя это может быть не так

Нулевая вероятность: поправка Лапласа

- Предположение
  - Обучающая выборка настолько большая, что добавление нескольких объектов не повлияет на вычисление $P(X|C_i)$
- $\exists x_j \{X \in C_i | X.A_j = x_j\} = \emptyset \Rightarrow$

$$P(X|C_i) = \prod_{j=1}^{n} \frac{|\{X \in C_i | X.A_j = x_j\}|}{|C_i|} \approx \prod_{j=1}^{n} \frac{|\{X \in C_i | X.A_j = x_j\}| + \mathbf{1}}{|C_i| + \mathbf{|A_j|}}$$

Пьер-Симон де Лаплас
1749-1827

- Пример
  - Пусть в обучающей выборке 1000 здоровых пациентов: 990 имеют нормальный гемоглобин и 10 – высокий, 0 – низкий
  - Тогда
    - $P(\text{Низкий}|\text{Здоров}) = \frac{0+1}{1000+(1+1+1)} = 0.001$
    - $P(\text{Норма}|\text{Здоров}) = \frac{990+1}{1000+(1+1+1)} = 0.988$
    - $P(\text{Высокий}|\text{Здоров}) = \frac{10+1}{1000+(1+1+1)} = 0.011$

Проблема некатегориальных атрибутов

- Если атрибут A_j – непрерывный, а не дискретный, то нельзя вычислить

$$P(X|C_i) = \prod_{j=1}^{n} \frac{|\{X \in C_i | X.A_j = x_j\}|}{|C_i|}$$

- Полагаем, что атрибут имеет нормальное (Гауссово) распределение

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- $P(A_j = x_j | C_i) = g(x_j, \mu_{C_i}, \sigma_{C_i})$

Карл Фридрих Гаусс
1777-1855

Гауссово (нормальное) распределение

- $P(A_j = x_j | C_i) = g(x_j, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{-\frac{(x_j - \mu_{C_i})^2}{2\sigma_{C_i}^2}}$
  - Среднее значение по атрибуту A_j среди объектов класса C_i

$$\mu_{C_i} = \frac{1}{|C_i|} \sum_{k=1}^{|C_i|} A_{j_k}$$

  - Стандартное отклонение по атрибуту A_j среди объектов класса C_i

$$\sigma_{C_i} = \sqrt{\frac{1}{|C_i|} \sum_{k=1}^{|C_i|} A_{j_k}^2 - \mu_{C_i}^2}$$

- Пример
  - Пусть *Возраст* – непрерывный атрибут, здоровые пациенты имеют возраст 36..60 лет. Нужно найти $P(\text{Возраст} = 45 | \text{Здоров})$.
  - Тогда в формулу выше подставляем $x_j = 45$, $\mu = 48$, $\sigma = 12$.

Почему **наивная** Байесовская классификация?

- $P(X|C_i) = \prod_{j=1}^{n} P(A_j = x_j|C_i) \Leftrightarrow$
  атрибуты $A_1, \ldots, A_n$ независимы друг от друга
- Но на практике это часто не так: атрибуты влияют друг на друга
  - Пример. У больных пациентов высокая температура часто наблюдается совместно с кашлем и низким гемоглобином: атрибуты t^0, Кашель, Гемоглобин не являются независимыми

Байесовская классификация: за и против

- Преимущества
  - Простая реализация
  - Точность классификации, сравнимая с деревьями решений и нейронными сетями в некоторых предметных областях
  - Используется как baseline (теоретическое обоснование для других классификаторов, которые явно не используют теорему Байеса)
- Недостатки
  - Потеря точности в случае зависимых друг от друга атрибутов

Литература

- Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd Edition. Morgan Kaufmann, 2011. 740 p. ISBN 978-0123814791
  - 8.3. Bayes Classification Methods, pp. 350-355
- Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. 1st Edition. Pearson, 2019. 732 p. ISBN: 978-1-292-02615-2
  - 5.3 Bayesian Classifiers, pp. 227-240