Введение в АД

Лекция 3

Машинное обучение

Искусственный интеллект (AI)

общее направление в компьютерных науках, включает все методы и технологии, для имитации человеческий интеллект.

Машинное обучение (ML)

один из подходов к технологической реализации AI.

Обучение с учителем (Supervised learning)

раздел ML, в котором модели обучаются на размеченных данных.

Блок занятий по ML в курсе ВвАД

1. Линейные модели (*сегодня*)
2. Нейронные сети
3. Компьютерное зрение (CV)
4. Обработка естественного языка (NLP)
5. Кластеризация (*обучение без учителя*)

Обучение с учителем

$X_1, ..., X_n$ — объекты из множества $\mathscr{X}$

$Y_1, ..., Y_n$ — таргеты из множества $\mathscr{Y}$

Требуется подобрать функцию $y : \mathscr{X} \to \mathscr{Y}$,
приближающую исходную зависимость.

Обучение с учителем

$X_1, ..., X_n$ — объекты из множества $\mathscr{X}$

$Y_1, ..., Y_n$ — таргеты из множества $\mathscr{Y}$

Регрессия

$\mathscr{Y}$ — $\mathbb{R}$, $\mathbb{R}_+$ или интервал в $\mathbb{R}$,

Примеры регрессии

- *Прогнозирование спроса* – предсказание количества проданных товаров в следующем месяце.
- *Предсказание уровня сахара в крови* – моделирование изменений глюкозы у диабетиков в зависимости от питания и активности.

Классификация

$\mathscr{Y}$ — конечное множество

Примеры классификации

- *Распознавание спама* – определение, является ли письмо спамом или нет.
- *Диагностика заболеваний* – определение наличия или отсутствия болезни по снимкам МРТ
- *Анализ тональности текста* – определение, является ли отзыв положительным

Линейная регрессия

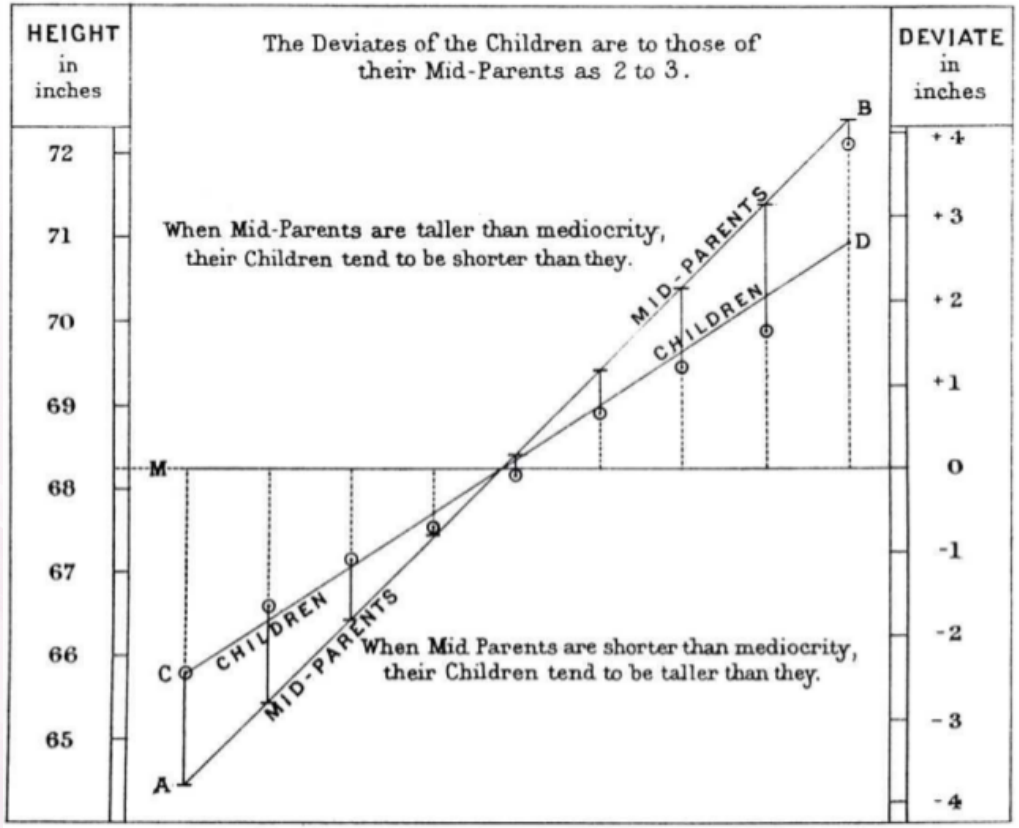
Первое упоминание регрессии

Впервые регрессия упоминается в работе Гальтона

"*Регрессия к середине в наследственности роста*", 1885 г.

Пусть x — рост родителей, y — рост детей.

Установлена зависимость $y - \overline{y} \approx \frac{2}{3}(x - \overline{x})$, т.е. регрессия к середине.



Модель линейной регрессии

Пусть $\mathscr{X} \subset \mathbb{R}^d$ — множество признаков, $\mathscr{Y} = \mathbb{R}$ — таргеты.

Рассматриваем зависимость вида

$$y(x) = \theta_1 x_1 + ... + \theta_d x_d,$$

где $x_1, ..., x_d$ — признаки,

$\theta = (\theta_1, ..., \theta_d)^T$ — вектор параметров.

Простой пример

$y = \theta_0 + \theta_1 x$,

x — рост котика,

y — потребление еды,

θ_0, θ_1 — неизвестные параметры.

Зависимость

- **линейна по параметрам**,
- линейна по аргументу.

Более сложный пример

$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2$,

x_1 — рост котика,

x_2 — вес котика,

y — потребление еды,

$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры.

Зависимость

- **линейна по параметрам**,
- квадратична по аргументам.

Нелинейные признаки

Зависимость $y = y(x)$ должна быть **линейна по параметрам**, но не обязана быть линейной по признакам.

Пусть $z_1, ..., z_k$ — набор "независимых" переменных. Можно рассматривать модель $y(x) = \theta_1 x_1(z_1, ..., z_k) + ... + \theta_d x_d(z_1, ..., z_k)$, где $x_j(z_1, ..., z_k)$ — некоторые функции, м.б. нелинейные.

Примеры: $x(z_1, ..., z_k) = 1$, $x(z_1, ..., z_k) = z_1$, $x(z_1, ..., z_k) = z_1^2 \ln z_2$.

Матричная запись

Представим данные в матричном виде

$$Y = \begin{pmatrix} Y_1 \\ ... \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & ... & x_{1d} \\ ... & & \\ x_{n1} & ... & x_{nd} \end{pmatrix}.$$

Линейная регрессия предполагает зависимость $Y = X\theta$.

Пример: Потребление мороженого

Предполагается линейная зависимость потребления мороженого в литрах на человека от среднесуточной температуры воздуха: $ic = \theta_0 + \theta_1 t$.

В этом примере $x_0(t) = 1, \quad x_1(t) = t,$

$$X = \begin{pmatrix} 1 & t_1 \\ \dots & \\ 1 & t_n \end{pmatrix}, \; Y = \begin{pmatrix} IC_1 \\ \dots \\ IC_n \end{pmatrix}, \; \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}.$$

Пусть $w = I\{\text{выходной день}\}$, зависимость $ic = \theta_0 + \theta_1 t + \theta_2 t^2 w$.
В этом примере $x_0(t, w) = 1, \quad x_1(t, w) = t, \quad x_2(t, w) = t^2 w,$

$$X = \begin{pmatrix} 1 & t_1 & t_1^2 w_1 \\ \dots & & \\ 1 & t_n & t_n^2 w_n \end{pmatrix}, \; Y = \begin{pmatrix} IC_1 \\ \dots \\ IC_n \end{pmatrix}, \; \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}.$$

Материал по доске

Отступление в матричное дифференцирование

Пусть $f : \mathbb{R}^n \to \mathbb{R}$. Тогда

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f}{\partial x_1} & \dots & \frac{\partial f}{\partial x_n} \end{pmatrix}$$ — производная (вектор-строка)

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$ — градиент (вектор-столбец).

Пример 1

$f(x) = a^T x$, где $a, x \in \mathbb{R}^n$

$$\frac{\partial f}{\partial x} = a^T$$ — производная

$$\nabla f = a$$ — градиент

Пример 2

$f(x) = x^T A x$, где $x \in \mathbb{R}^n$,

матрица $A \in \mathbb{R}^{n \times n}$ симметрична

$$\frac{\partial f}{\partial x} = 2x^T A$$ — производная

$$\nabla f = 2Ax$$ — градиент

Пример

Данные:

x	0	1	2
y	0	4	7

Модель

$$y(x) = \theta_0 + \theta_1 x$$

Найдите оценку коэффициентов.

Решение

Матричный вид: $X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$, $Y = \begin{pmatrix} 0 \\ 4 \\ 7 \end{pmatrix}$, $\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}$.

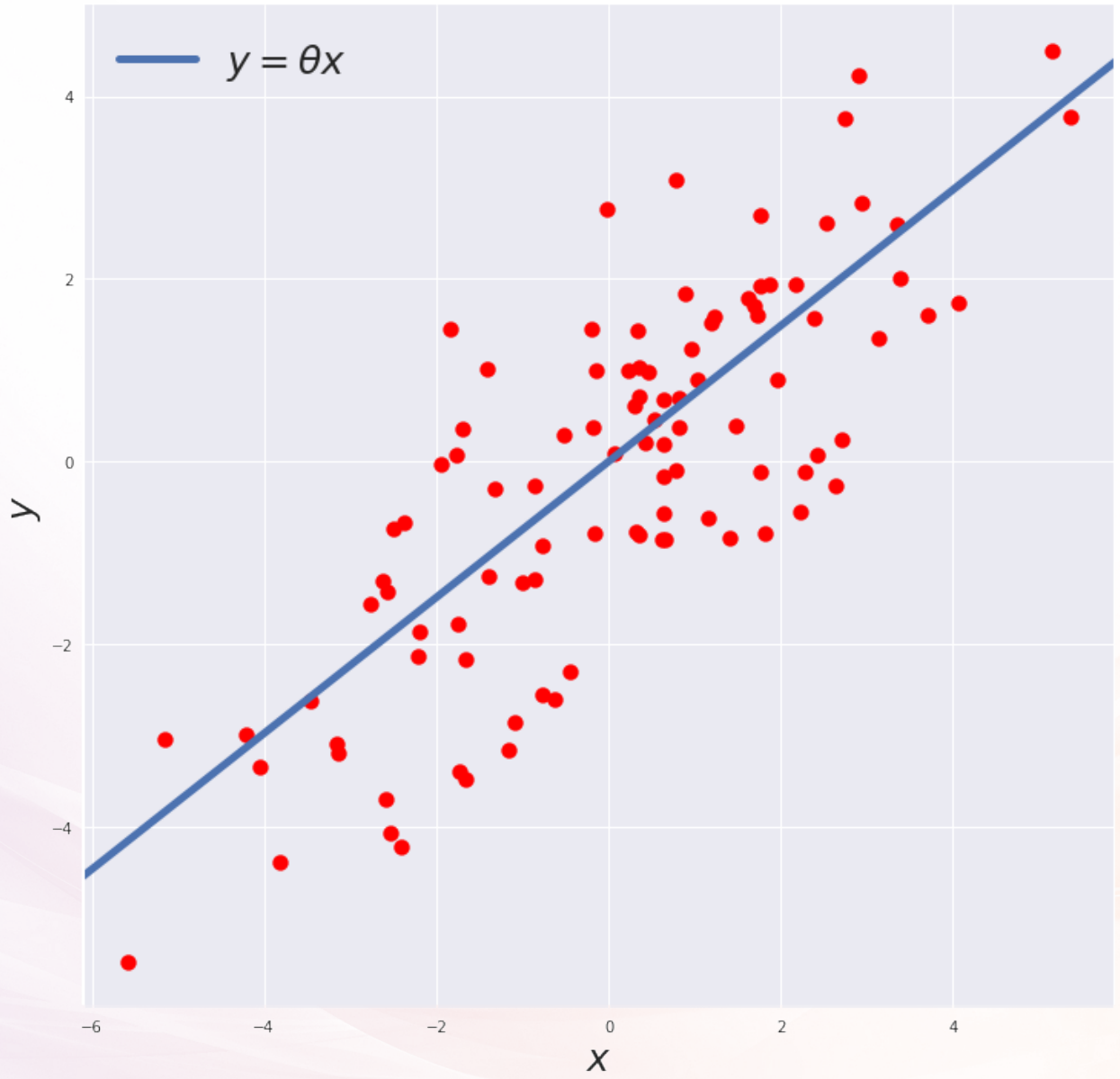
Оценка коэффициентов $\widehat{\theta} = (X^T X)^{-1} X^T Y = \begin{pmatrix} 1/6 \\ 21/6 \end{pmatrix}$

Обученная модель

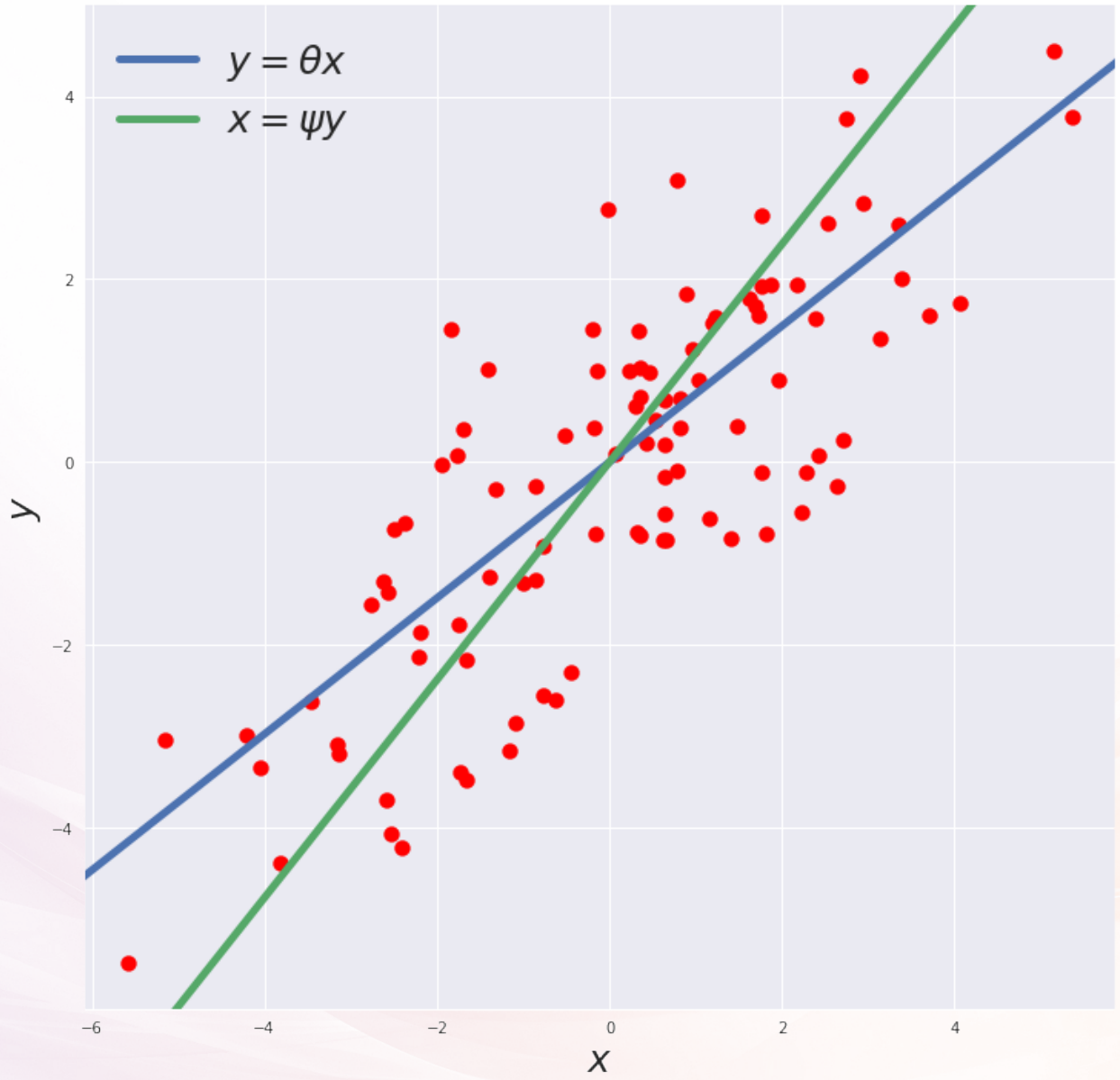
$$y(x) = \frac{1}{6} + \frac{21}{6} x$$

Материал по доске

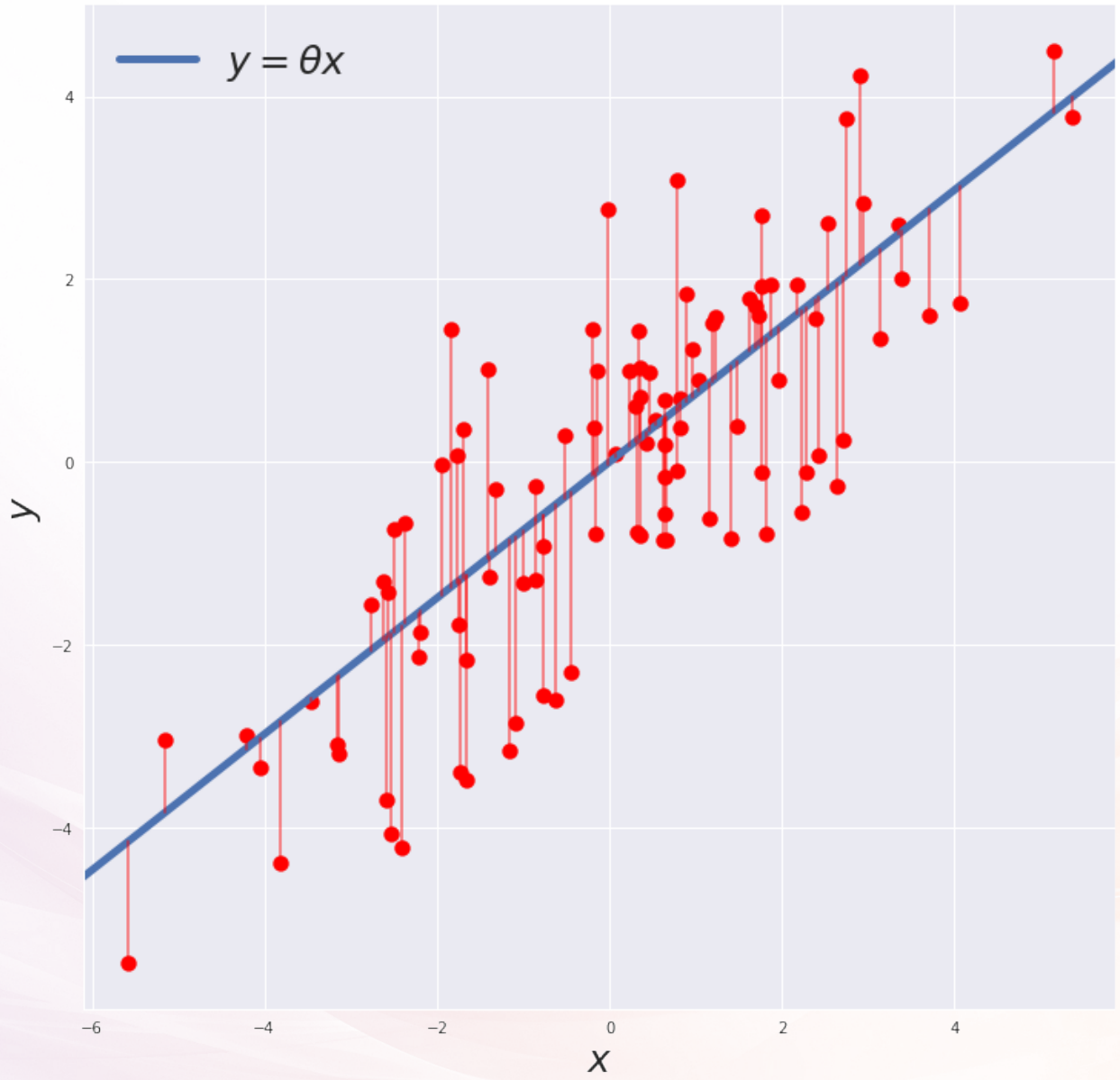
Инверсия



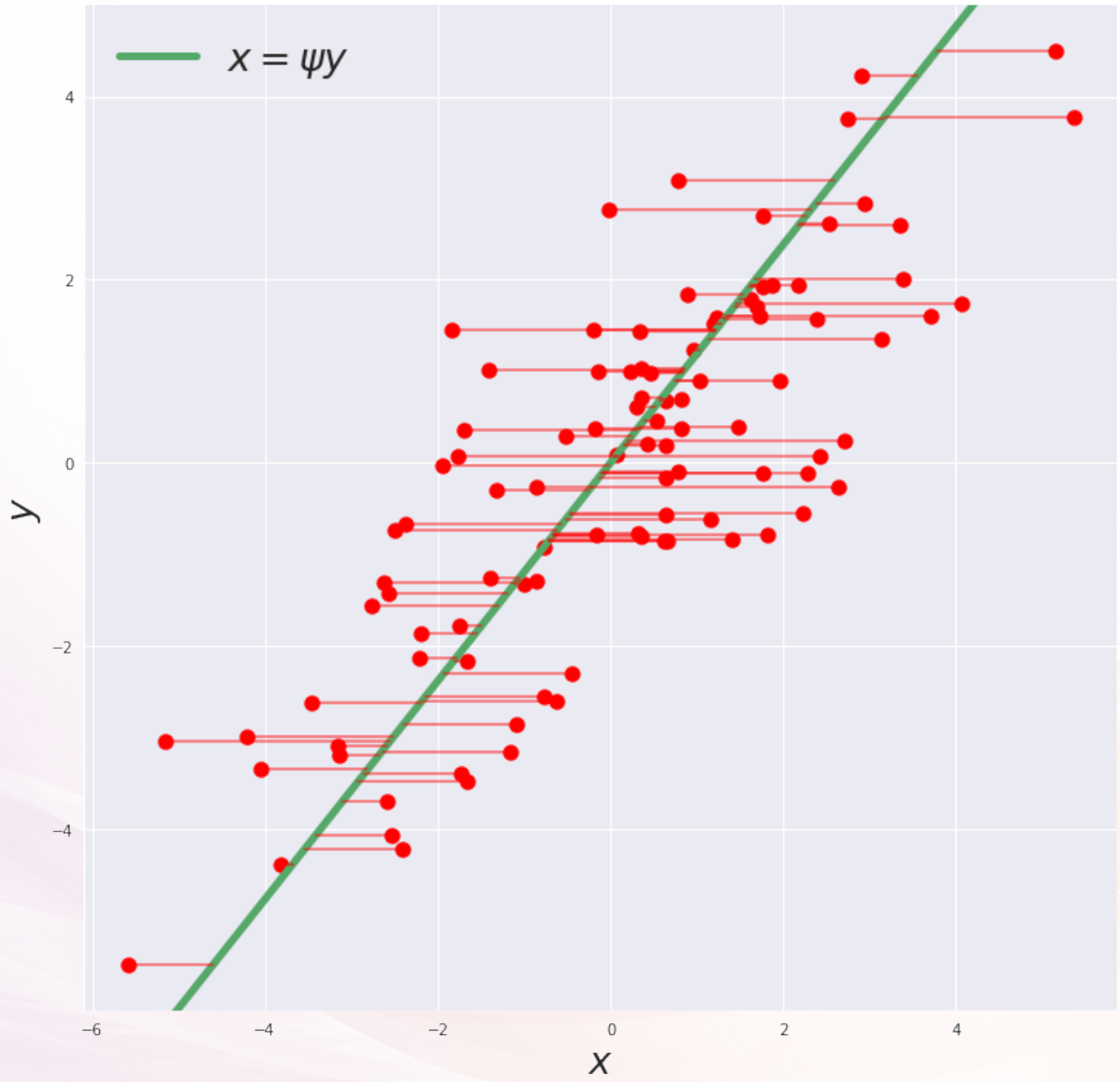
Инверсия



Инверсия



Инверсия



Задача классификации

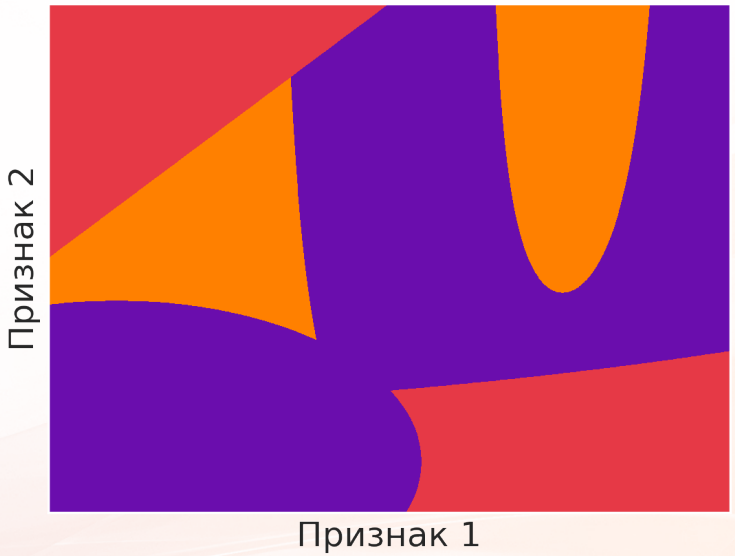
Классификация

$\mathscr{X}$ — пространство объектов,

$\mathscr{Y}$ — конечное множество классов.

Правило классификации: $y : \mathscr{X} \to \mathscr{Y}$.

Пространство $\mathscr{X}$ разбивается на *подпространства (decision regions)* $\mathscr{X}_k = \{x \in \mathscr{X} \mid y(x) = k\}$, границы которых называются *разделяющими поверхностями (decision surfaces)*.



Вероятностная природа

Часто предполагается случайная принадлежность к классу:

при повторении эксперимента один и тот же объект $x \in \mathscr{X}$ может быть отнесен как одному классу, так и к другому.

$\Longrightarrow$ имеет смысл **предсказывать вероятность** $\mathrm{P}_x(Y = k)$ принадлежности объекта x каждому из классов.

Точечная оценка: $\arg\max\limits_{k \in \mathscr{Y}} \mathrm{P}_x(Y = k)$

Если классы неравнозначны:
$\arg\max\limits_{k \in \mathscr{Y}} [w_k \, \mathrm{P}_x(Y = k)]$,
w_k — приоритетность класса

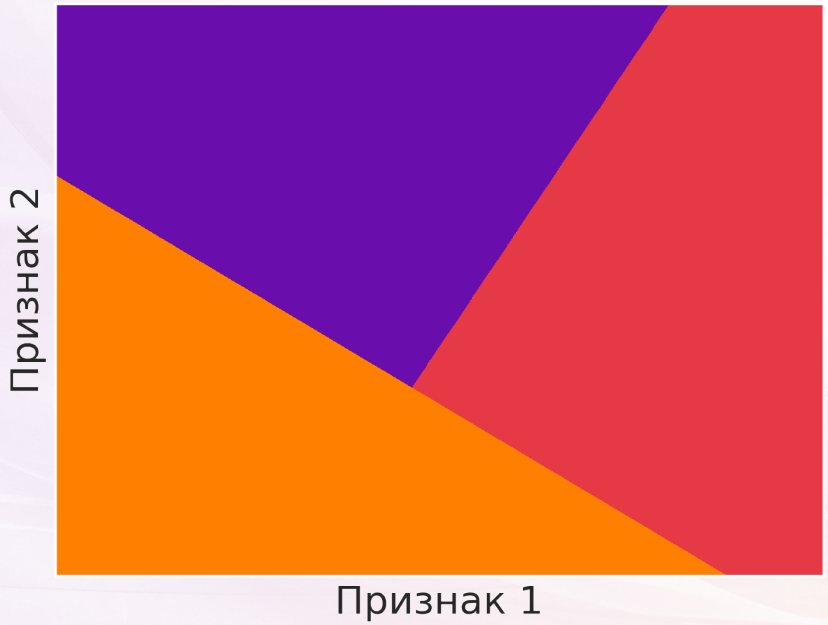
Примеры:

1. $\mathrm{P}(Y = 0 \mid X = x_2) = 0.95, \quad \mathrm{P}(Y = 1 \mid X = x_2) = 0.05$
   уверенное предсказание в пользу класса 0
2. $\mathrm{P}(Y = 0 \mid X = x_1) = 0.55, \quad \mathrm{P}(Y = 1 \mid X = x_1) = 0.45$
   модель не уверена в предсказании.

Линейные модели

$y(x) = \theta^T x$ — линейная модель регрессии.

Модель классификации называется линейной если разделяющая поверхность — линейная гиперплоскость в пр-ве $\mathscr{X}$.
В многоклассовом случае — при дополнении до гиперплоскости.
Пример: при $\mathscr{Y} = \{0, 1\}$ *линейна модель* $y(x) = \mathrm{sign}(\theta^T x)$.



Замечание.
Исходное пространство признаков может быть предварительно преобразовано с помощью нелинейных функций, в частности можно включить константный признак. В таком случае разделяющая поверхность линейного классификатора не будет линейной в исходном пространстве.

Логистическая регрессия

Материал по доске

Отступление
в теорию информации

Кодирование

Алфавит: {A, B, C}

Как его закодировать с помощью 0 и 1?

Правило кодирования:

A $\longrightarrow$ 00
B $\longrightarrow$ 01
C $\longrightarrow$ 10

Дано сообщение:

AAAAAACAABAAACAAAAACABAAA
AAAAAABABCBABAAABBAABAAAA

Закодированное сообщение:

00 00 00 00 00 00 10 00 00 01 00 00 00 10 00 00 00 00 00 10 00 01 00 00 00
00 00 00 00 00 00 01 00 01 10 01 00 01 00 00 00 01 01 00 00 01 00 00 00 00

Длина символа: 2

Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):

Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):

A	B	C
вер. = 0.7	вер. = 0.2	вер. = 0.1

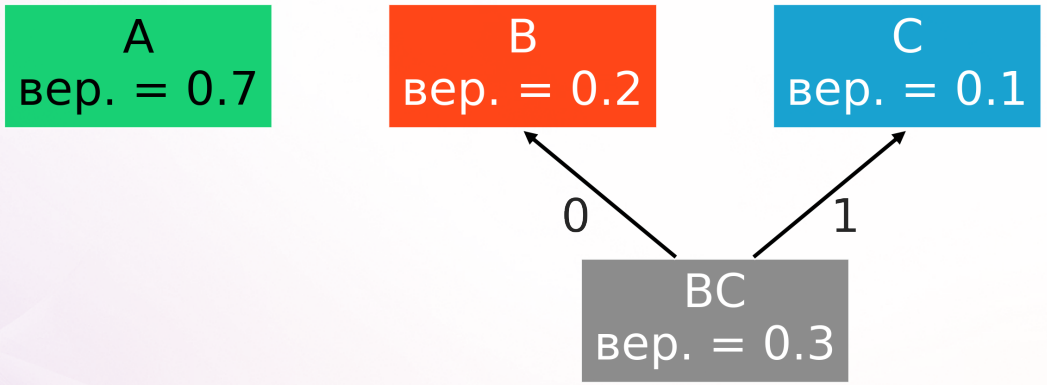
Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):



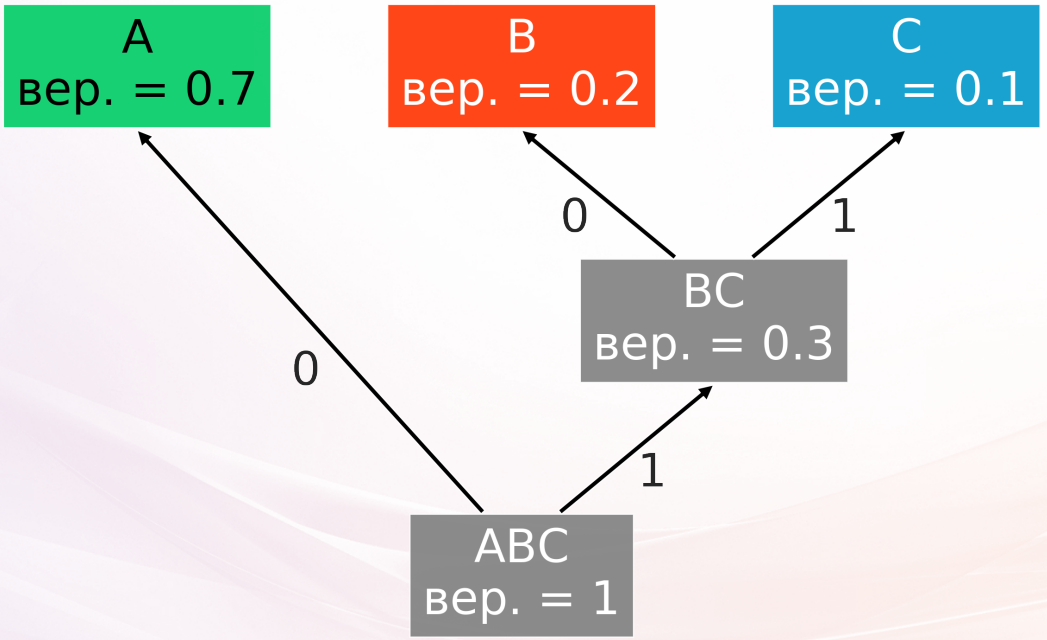
Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):



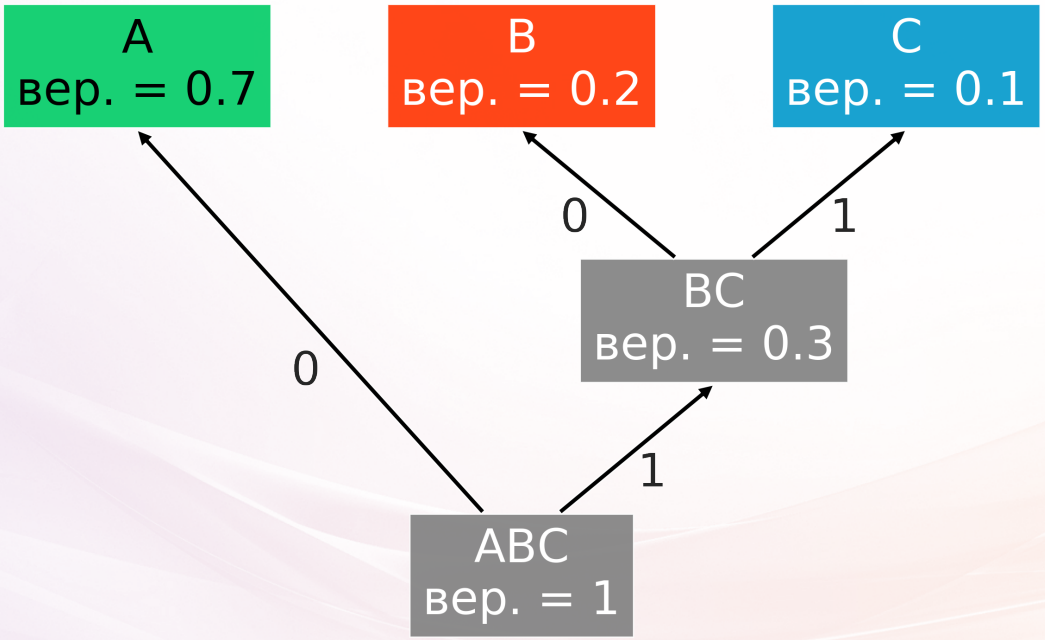
Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):



Правило кодирования:

A $\longrightarrow$ **0**

B $\longrightarrow$ **10**

C $\longrightarrow$ **11**

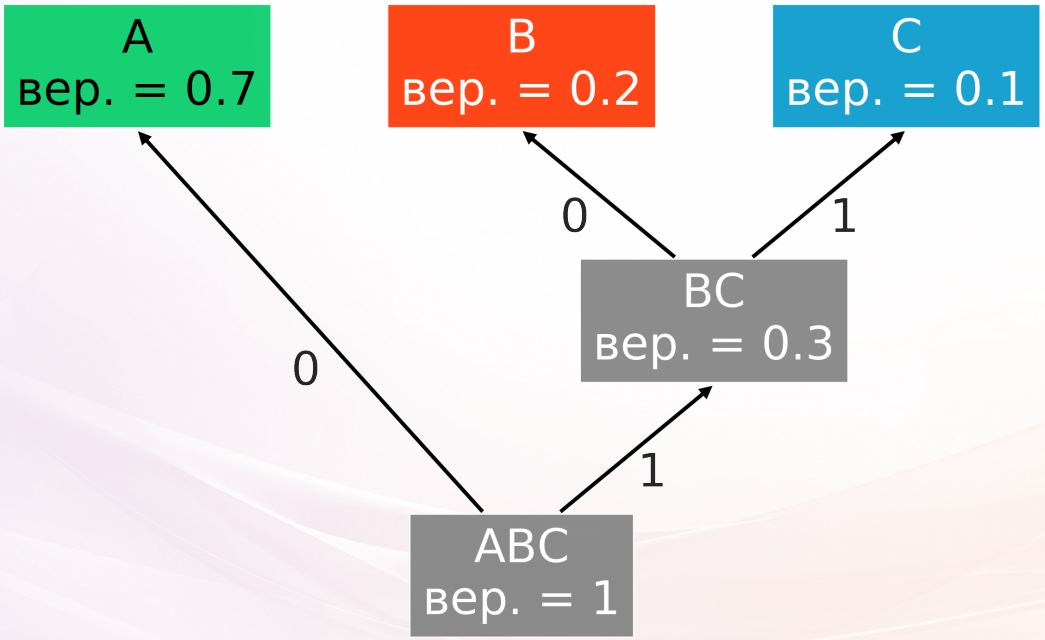
Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):



Правило кодирования:

A $\longrightarrow$ **0**

B $\longrightarrow$ **10**

C $\longrightarrow$ **11**

Декод-ние однозначно, т.к. ни один код не является префиксом другого.

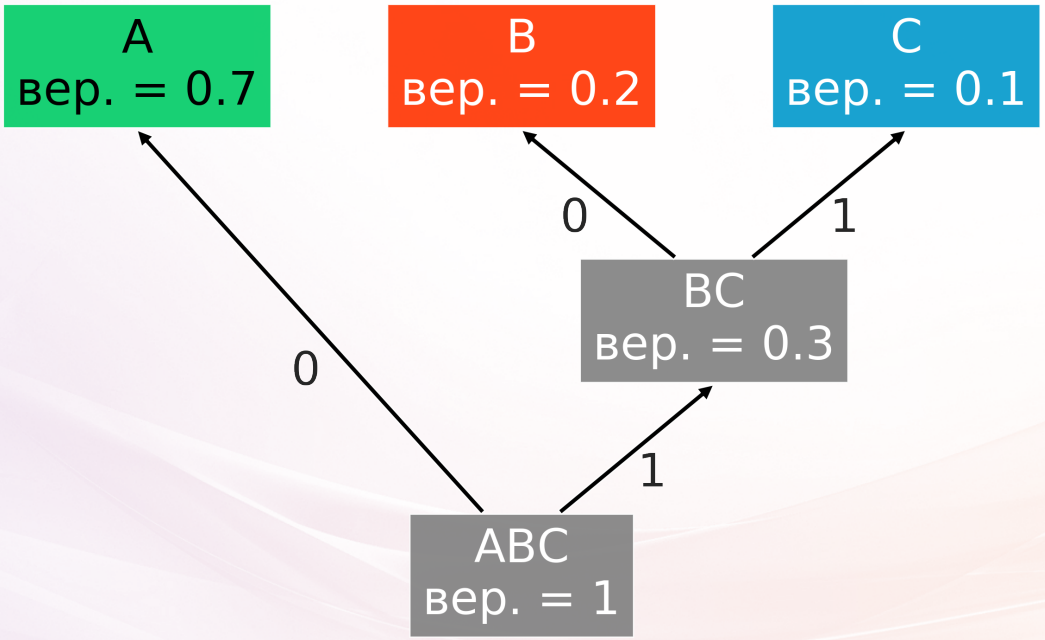
Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):



Правило кодирования:

A $\longrightarrow$ **0**

B $\longrightarrow$ **10**

C $\longrightarrow$ **11**

Декод-ние однозначно, т.к. ни один код не является префиксом другого.

Средняя длина символа:

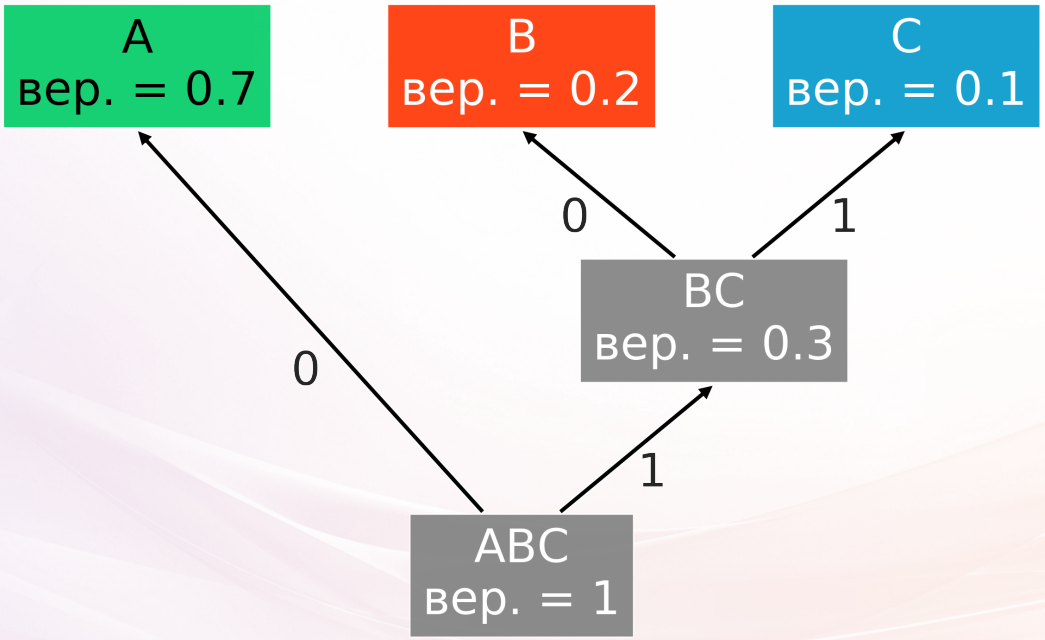
Код Хаффмана

Алфавит: {**A**, **B**, **C**}

Известны вероятности появления символов: {**0.7**, **0.2**, **0.1**}

Хочется уменьшить длину закодированного сообщения.

Метод построения оптимального кода (Хаффман):



Правило кодирования:

A $\longrightarrow$ **0**

B $\longrightarrow$ **10**

C $\longrightarrow$ **11**

Декод-ние однозначно, т.к. ни один код не является префиксом другого.

Средняя длина символа: $0.7 \cdot 1 + 0.2 \cdot 2 + 0.1 \cdot 2 = 1.3$

Пример

Исходное сообщение

AAAAAACAABAAACAAAAACABAAA
AAAAAABABCBABAAABBAABAAAA

Первое кодирование

00 00 00 00 00 00 10 00 00 01 00 00 00 10 00 00 00 00 00 10 00 01 00 00 00
00 00 00 00 00 00 01 00 01 10 01 00 01 00 00 00 01 01 00 00 01 00 00 00 00

Количество символов: 100; Длина символа: 2

Код Хаффмана

Правило кодирования:

A $\longrightarrow$ 0
B $\longrightarrow$ 10
C $\longrightarrow$ 11

Закодированное сообщение:

0 0 0 0 0 0 11 0 0 10 0 0 0 11 0 0 0 0 0 11 0 10 0 0 0 0 0 0 0 0 0 10 0 10 11 10
0 10 0 0 0 10 10 0 0 10 0 0 0 0

Количество символов: 63; Средняя длина символа: 1.3

Средняя длина символа

Утверждение:

Для кодирования символа, встречающегося с вероятностью p_j в "идеале" нужно $\log_2 \frac{1}{p_j}$ бит. Приближение — коды Хаффмана.

Пример:

Символы равновероятны $\Rightarrow$ для каждого символа нужно $\lceil \log_2 k \rceil$ бит.

Пусть символы $a_1, ..., a_k$ встречаются с вер-тями $p_1, ..., p_k$.

Энтропия — средняя длина символа при оптимальном кодировании.

$$H(\mathrm{P}) = -\sum_{j=1}^{k} p_j \log_2 p_j$$

В нашем примере для вероятностей $\{\mathbf{0.7}, \mathbf{0.2}, \mathbf{0.1}\}$

$H(\mathrm{P}) = -0.7 \log_2 0.7 - 0.2 \log_2 0.2 - 0.1 \log_2 0.1 \approx 1.157$

А мы построили код со средней длинной символа 1.3.

Кодирование с помощью другого распределения

Что будет, если будем кодировать кодом, построенным по распр. Q $= \{q_1, ..., q_k\}$, если истинное распр. P $= \{p_1, ..., p_k\}$?

Алфавит: {**A**, **B**, **C**}

Истинные вероятности появления символов: P ={**0.7**, **0.2**, **0.1**}

Предполагаемые вер-ти появления символов: Q ={**0.4**, **0.5**, **0.1**}

Правило кодирования для Q:

A $\longrightarrow$ **10**
B $\longrightarrow$ **0**
C $\longrightarrow$ **11**

Закодированное сообщение:

10 10 10 10 10 10 11 10 10 0 10 10 10 11 10 10 10 10 10 11 10 0 10 10 10 10
10 10 10 10 10 0 10 0 11 0 10 0 10 10 10 0 0 10 10 0 10 10 10 10

Количество символов: 91

Средняя длина символа: $0.7 \cdot 2 + 0.2 \cdot 1 + 0.1 \cdot 2 = 1.8$

Кодирование с помощью другого распределения

Что будет, если будем кодировать кодом, построенным по распр. $\mathrm{Q} = \{q_1, ..., q_k\}$, если истинное распр. $\mathrm{P} = \{p_1, ..., p_k\}$?

Кросс-энтропия — средняя длина символа при кодировании алфавита вероятностями появления символов Q,
если на самом деле они появляются с вероятностями P.

$$H(\mathrm{P}, \mathrm{Q}) = -\sum_{j=1}^{k} p_j \log_2 q_j$$

Дивергенция Кульбака-Лейблера — избыточная длина символа при кодировании алфавита вероятностями появления символов Q, если на самом деле они появляются с вероятностями P.

$$KL(\mathrm{P}, \mathrm{Q}) = H(\mathrm{P}, \mathrm{Q}) - H(\mathrm{P}) = \sum_{j=1}^{k} p_j \log_2 \frac{p_j}{q_j}$$

Кодирование с помощью другого распределения

Алфавит: {**A**, **B**, **C**}

Истинные вероятности появления символов: P ={**0.7**, **0.2**, **0.1**}

Предполагаемые вер-ти появления символов: Q ={**0.4**, **0.5**, **0.1**}

$$H(\mathrm{P}) = -0.7 \log_2 0.7 - 0.2 \log_2 0.2 - 0.1 \log_2 0.1 \approx 1.157$$

$$H(\mathrm{P}, \mathrm{Q}) = -0.7 \log_2 0.4 - 0.2 \log_2 0.5 - 0.1 \log_2 0.1 \approx 1.458$$

$$KL(\mathrm{P}, \mathrm{Q}) = H(\mathrm{P}, \mathrm{Q}) - H(\mathrm{P}) \approx 1.458 - 1.157 = 0.301$$

В теории мы тратим лишние 0.3 бита на символ.

Для приближающих кодов Хаффмана:

- Средняя длина символа при кодировании по P: 1.3
- Средняя длина символа при кодировании по Q: 1.8
- Избыточная длина символа: 0.5

ВСЁ!