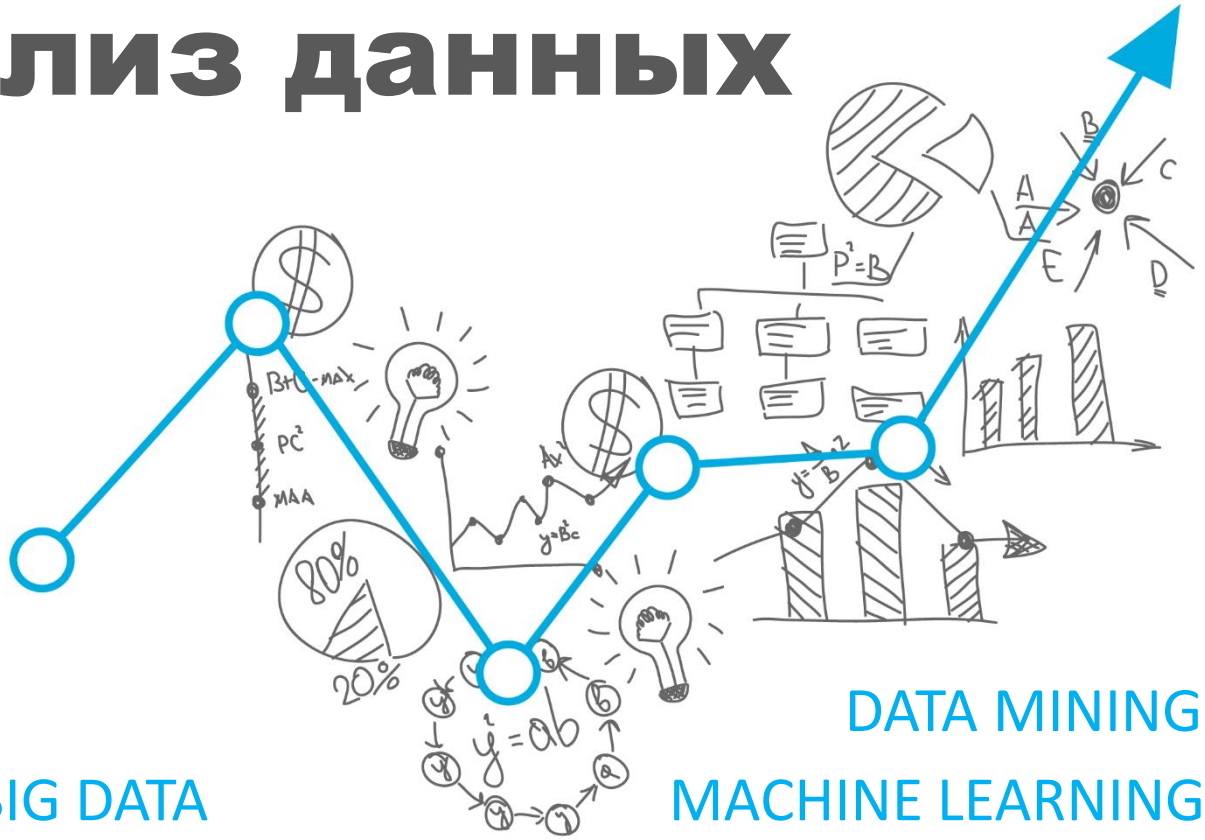
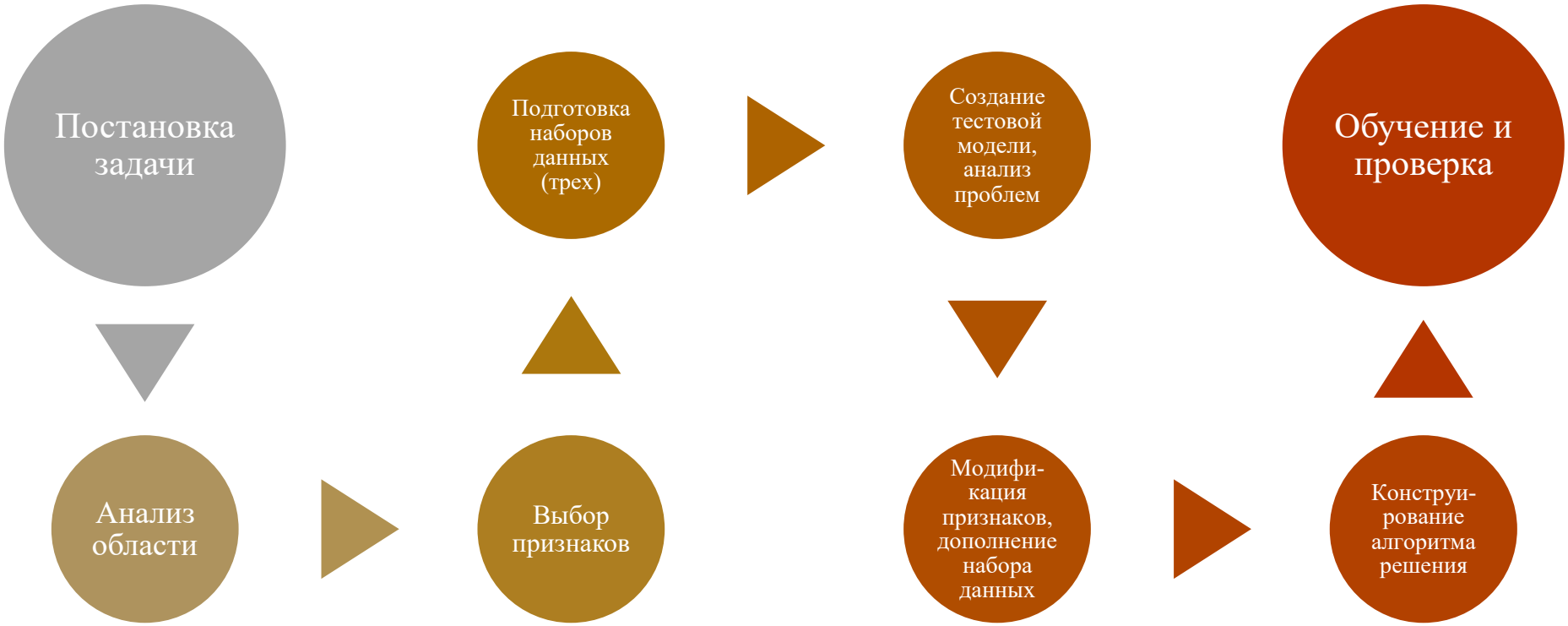
Интеллектуальный анализ данных



Лекция 8. Метрики машинного обучения

Проектирование машинного обучения

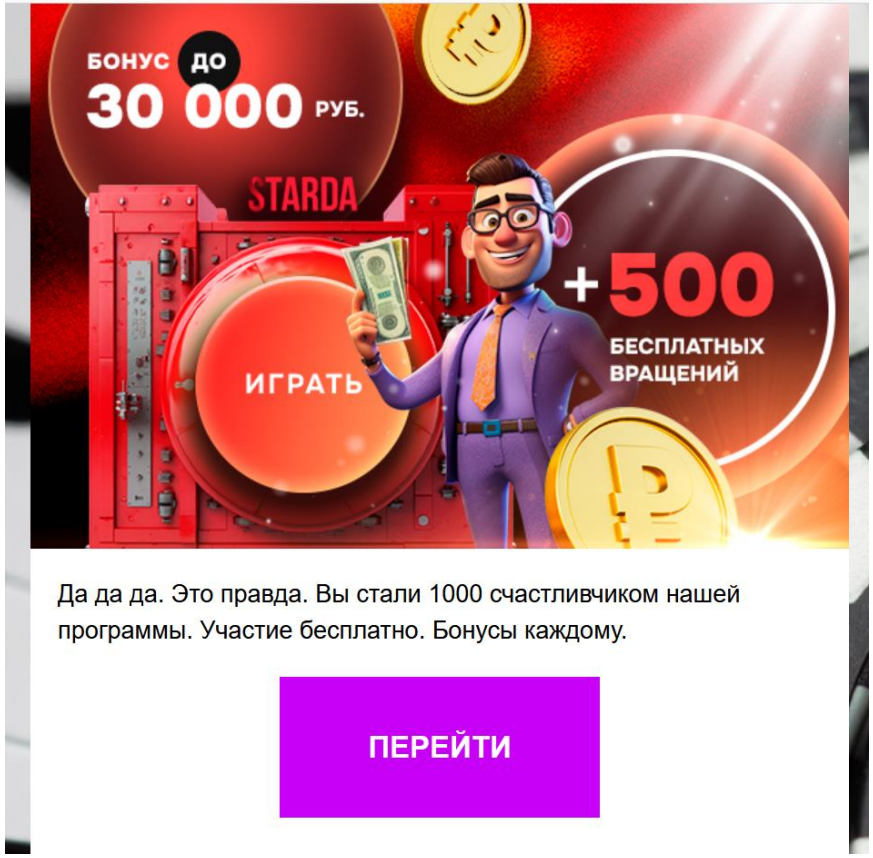


Пример: классификация спама

От: Пополнили счет <cijxwd@businesslife-online.ru>

Кому: cdg@mail.ru

Тема: Уведомление о зачислении 30 000 руб



Спам

От: Пополнили счет <cijxwd@businesslife-online.ru>

Кому: cdg@mail.ru

Тема: Уведомление о зачислении 30 000 руб

Уважаемые члены ученого совета! Очередное заседание №6 состоится 16.01.2024 в 14.00 в 305 лк. Прошу обратить внимание на нетрадиционное время заседания (в связи с сессией)

Повестка заседания:
1. Эффективные контракты с НПС – анализ результативности в 2022-2023 гг.(вопрос прошлого заседания) (Гринченков)
2. Об итогах второй аттестации и ходе зимней сессии (Саенко)
3. Утверждение дополнительных программ к программе-минимум кандидатского экзамена
4. Конкурс (каф. ИИСТ)

Ученый секретарь Кузнецова А.В.

ФГБОУ ВО "ЮРГПУ (НПИ) имени М.И.Платова"

Не спам

Пример: классификация спама

Классы $y \in \{0, 1\}$ $y = 0$ – не спам

$y = 1$ – спам

Признаки x ?

Выбрать 100 слов: распродажа, каталог, скидка, промокод, акция, предложение, сейчас, покупай, …

От: mail@mosmexa.ru

Кому: cdg@mail.ru

Тема: Тотальная распродажа новой коллекции!

Московская меховая компания **проводит январскую распродажу Новой коллекции** пуховиков! В нашем каталоге более **260** женских и мужских моделей на разный вкус. Размеры - от **40** до **70**, более **10** цветовых решений! И приятный бонус - **до 21 января** вы можете получить скидку в размере **30%** на акционные пуховики, введя в корзине промокод **20240113**.

$$x = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ \vdots \\ \\ \end{bmatrix} \begin{matrix} \text{покупай} \\ \text{распродажа} \\ \text{скидка} \\ \text{каталог} \\ \text{сейчас} \\ \vdots \end{matrix}$$

$x \in \mathbb{R}^{100}$

Реальные системы:

$x \in \mathbb{R}^{50\,000 \ldots 100\,000}$

Пример: классификация спама

Как обучить этот классификатор и добиться минимальной ошибки?

- Собрать много данных для обучения и контроля

  Сбор писем (honeypot), использование готовых БД спама

- Разработать более сложные признаки и алгоритмы из получения

  Учитывающие источник письма, маршрут доставки и т.д.

  Учитывающие склонения, множественное число и т.д.

  скидка / скидки выгода / выгодное

  Разработать алгоритмы, учитывающие намеренные опечатки маскировки слов

  эксклюзивное эксклюзвное эксклюЗивн0е эксклюзивное

  Учитывающие пунктуацию

  !!!Скидка!!!

Тестовая модель

- Рекомендуется начать с простой быстро реализуемой модели, которая позволит получить сразу какой-то ответ, который можно проанализировать.

  Заранее сложно выбрать сразу хорошее решение. Тестовая модель поможет лучше понять задачу.

- Построить кривые обучения, оценить требуется ли больше данных или больше признаков и т.д.

  Даже простая модель даст больше представления о особенностях задачи, чем безосновательная интуиция.

- Выполнить **анализ ошибок** тестовой модели

  Оценить на каких классах больше ошибок, проанализировать причины. Возможно нужны признаки, учитывающие особенности этих классов, возможно нужно больше данных именно этих классов и т.д.

Пример: классификация спама

Используем валидационный набор, $m_{CV} = 500$

Тестовая модель допустила 100 ошибок

Вручную рассматриваем эти ошибки. Например, к каким типам писем они относятся: реклама, вымогательство, кража личной информации и т.д.

Реклама товаров: 5

Приглашения на сайты: 11

Попытки украсть пароль: 42

Прочие: 36

Сосредоточиться на поиске особенностей этого типа

Пробуем варианты признаков:

- Выделить наличие ошибок в словах
- Необычный маршрут доставки письма
- Оценивать однокоренные слова как один признак

Пробуем разные варианты и оцениваем их результативность на валидационной выборке

Почему рекомендуется выполнять анализ ошибок и оценивать эффективность применения различных вариантов решения на валидационном наборе по ошибке $J_{cv}(\theta)$, а не на тестовом наборе и ошибке $J_{test}(\theta)$?

- ☒ Валидационный набор обычно больше тестового
- ☒ Ошибка на валидационном наборе обычно меньше
- ☑ Если мы подберем вариант решения по тестовому набору, то не сможем оценить обобщающую способность модели
- ☒ Использование валидационного набора даст меньший набор признаков

Пример: классификация спама

Используем валидационный набор, $m_{CV} = 500$

Тестовая модель допустила 100 ошибок, т.е. **20% ошибок**

Реализуем вариант, проверяем на валидационном наборе, оцениваем эффект:

➢ Выделить наличие ошибок в словах	**15% ошибок**	**Принимается**
➢ Необычный маршрут доставки письма	**12% ошибок**	**Принимается**
➢ Оценивать однокоренные слова как один признак	**7% ошибок**	**Принимается**
➢ Не учитывать регистр слов	**9% ошибок**	**Отвергается**

Несимметричные классы

Рассмотрим задачу медицинской диагностики ракового заболевания:

$y = 0$ – пациент не имеет заболевания

$y = 1$ – пациент имеет рак

Построена модель, которая имеет **1%** ошибок Достоверность $A = 0.99$

Достоверность (accuracy):

$$A = \frac{T}{m_{test}}$$

T – количество правильных предсказаний

Хороший ли это результат?

Представим, что в реальности 0.5% пациентов имеют рак.
Является ли $A = 0.99$ по-прежнему хорошим результатом?

Возьмем гипотезу

$h_\theta(x) = 0$

(алгоритм решения отсутствует)

$A = 0.995$

99.5%

0.5% ошибок

Какое решение лучше?

$A_1 = 0.986$

$A_2 = 0.992$

$A_3 = 0.995$

Несимметричные классы

Рассмотрим задачу медицинской диагностики ракового заболевания:

$y = 0$ – пациент не имеет заболевания

$y = 1$ – пациент имеет рак

Построена модель, которая имеет **1%** ошибок Достоверность $A = 0.99$

Достоверность (accuracy):

$$A = \frac{T}{m_{test}}$$

T – количество правильных предсказаний

Если количество объектов одного класса существенно меньше, чем объектов другого класса ($\#[y = 1] \ll \#[y = 0]$), такие классы называют **несимметричными**.

Достоверность является плохим вариантом для сравнения решения при несимметричных классах!

Матрица ошибок

$y = 1$ – редкий класс в несимметричной задаче

Предсказанные	Актуальные $y = 1$	Актуальные $y = 0$
$h(x) = 1$	**TP**	**FP**
$h(x) = 0$	**FN**	**TN**

$$m = TP + FP + TN + FN$$

Достоверность (accuracy):

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

TP (True Positive) – Истинно положительный

TN (True Negative) – Истинно отрицательный

Правильное предсказание

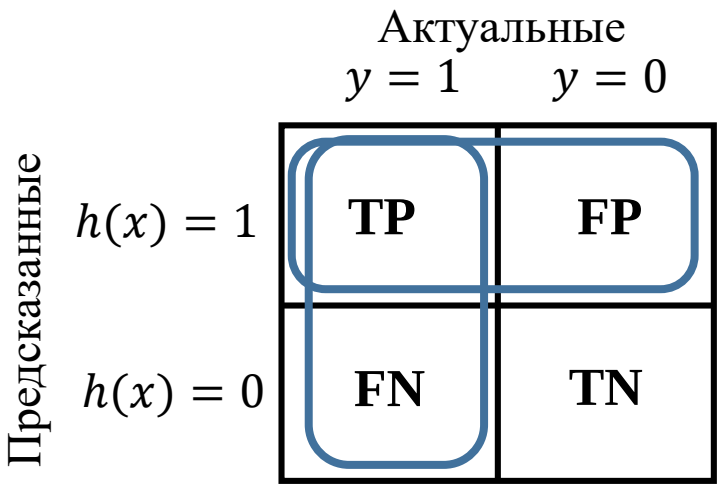
FP (False Positive) – Ложно положительный

FN (False Negative) – Ложно отрицательный

Ошибочное предсказание

Метрики точности и полноты

$y = 1$ – редкий класс в несимметричной задаче (только так, не наоборот!)



Точность (precision):

$$P = \frac{TP}{TP + FP}$$

Среди всех ответов $h(x) = 1$, какая часть совпадает с актуальными $y = 1$

Из всех пациентов, кому модель предсказала рак, какая часть действительно им больна

Полнота (recall):

$$R = \frac{TP}{TP + FN}$$

Среди всех актуальных $y = 1$, какая часть совпадает с ответами $h(x) = 1$

Из всех пациентов больных раком, кому модель правильно поставила диагноз

Возьмем гипотезу

$$h_\theta(x) = 0$$

$R = 0$ $P = 0$

Метрики точности и полноты

Предсказанные \ Актуальные	$y = 1$	$y = 0$
$h(x) = 1$	**80**	**20**
$h(x) = 0$	**80**	**820**

Достоверность (accuracy):

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

Точность (precision):

$$P = \frac{TP}{TP + FP}$$

Полнота (recall):

$$R = \frac{TP}{TP + FN}$$

Какова **точность** решения в соответствии с представленной таблицей ошибок?

- ☒ 0.08
- ☒ 0.50
- ☑ 0.80
- ☒ 0.90

Метрики точности и полноты

Предсказанные \ Актуальные	$y = 1$	$y = 0$
$h(x) = 1$	**80**	**20**
$h(x) = 0$	**80**	**820**

Достоверность (accuracy):

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

Точность (precision):

$$P = \frac{TP}{TP + FP}$$

Полнота (recall):

$$R = \frac{TP}{TP + FN}$$

Какова **полнота** решения в соответствии с представленной таблицей ошибок?

- ☒ 0.08
- ☑ 0.50
- ☒ 0.80
- ☒ 0.90

Метрики точности и полноты

Предсказанные \ Актуальные	$y = 1$	$y = 0$
$h(x) = 1$	**80**	**20**
$h(x) = 0$	**80**	**820**

Достоверность (accuracy):

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

Точность (precision):

$$P = \frac{TP}{TP + FP}$$

Полнота (recall):

$$R = \frac{TP}{TP + FN}$$

Какова **достоверность** решения в соответствии с представленной таблицей ошибок?

- ☒ 0.08
- ☒ 0.50
- ☒ 0.80
- ☑ 0.90

Метрики точности и полноты

Предсказанные \ Актуальные	$y = 1$	$y = 0$
$h(x) = 1$	**80**	**20**
$h(x) = 0$	**80**	**820**

Достоверность (accuracy):

$$\boldsymbol{A} = \frac{TP + TN}{TP + FP + TN + FN} = \mathbf{0.90}$$

Точность (precision):

$$\boldsymbol{P} = \frac{TP}{TP + FP} = \mathbf{0.80}$$

Полнота (recall):

$$\boldsymbol{R} = \frac{TP}{TP + FN} = \mathbf{0.50}$$

Какую метрику выбрать?

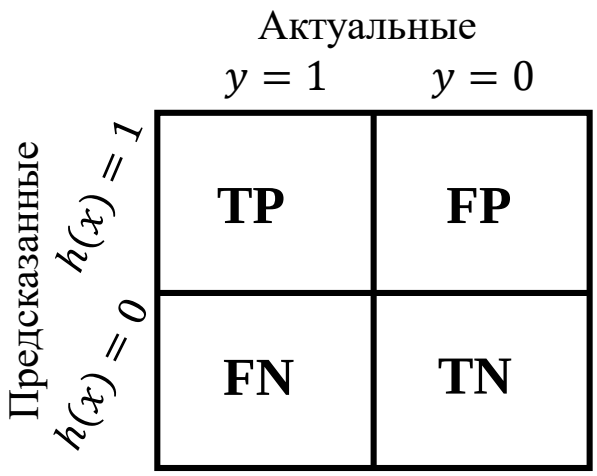
Достоверность не годится для несимметричных классов.
В чем разница между точностью и полнотой?

Компромисс между точностью и полнотой

Модель логистической регрессии: $0 \leq h(x) \leq 1$

Предсказываем 1, если $h(x) \geq$ ~~0.5~~ ~~0.7~~ 0.3

Предсказываем 0, если $h(x) <$ ~~0.5~~ ~~0.7~~ 0.3



1) Предположим, мы хотим говорить пациенту, что у него рак ($y = 1$), только если очень в этом уверены (избегать ложно-положительного результата)

Точность повышается, полнота уменьшается

2) Предположим, мы не хотим «пропустить» ни одного пациента с раком
(избегать ложно-отрицательного результата)

Точность понижается, полнота повышается

Точность (precision):

$$P = \frac{TP}{TP + FP}$$

Полнота (recall):

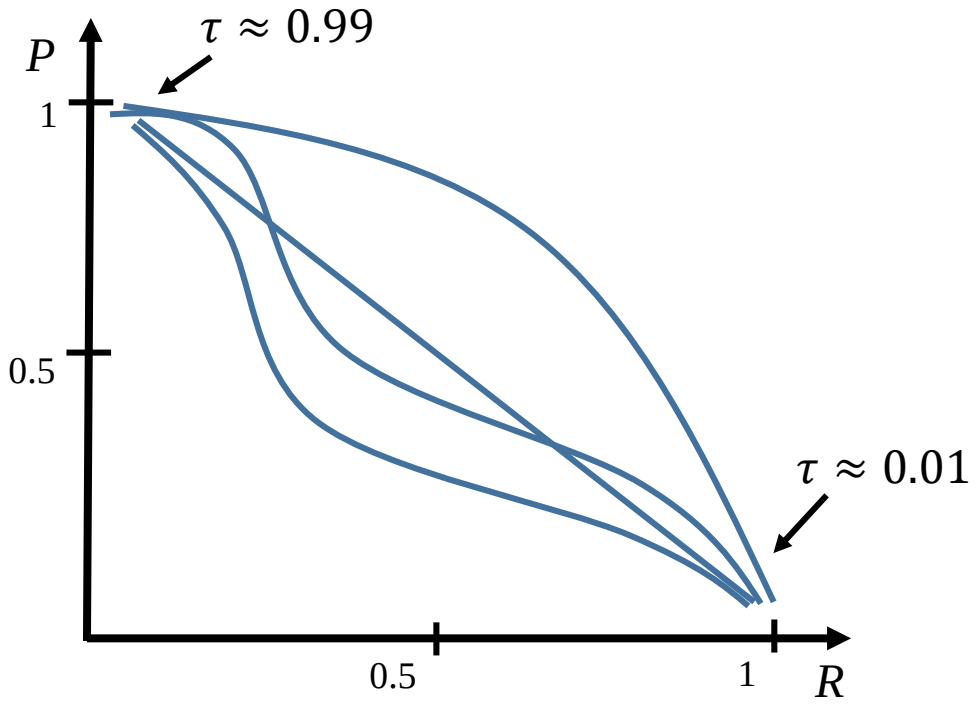
$$R = \frac{TP}{TP + FN}$$

Компромисс между точностью и полнотой

Модель логистической регрессии: $0 \le h(x) \le 1$

Предсказываем 1, если $h(x) \ge \tau$

Предсказываем 0, если $h(x) < \tau$



Предсказанные \ Актуальные	$y = 1$	$y = 0$
$h(x) = 1$	**TP**	**FP**
$h(x) = 0$	**FN**	**TN**

Точность (precision):

$$P = \frac{TP}{TP + FP}$$

Полнота (recall):

$$R = \frac{TP}{TP + FN}$$

Как обосновано выбрать значение τ ?

F-мера

Подобрать τ по валидационному набору, чтобы сочетание P и R было наилучшим

Сравнивать два число сложно, желательно из них получить одно число

Решение	Точность P	Полнота R	Среднее	F_1-мера
Алгоритм 1	0.5	0.4	0.45	0.44
Алгоритм 2	0.7	0.1	0.4	0.17
Алгоритм 3	0.02	1.0	0.51	0.04

лучший вариант

худший вариант

$h_\theta(x) = 1$

Среднее: $Avg = \frac{P + R}{2}$

$P = 0, R = 0 \quad \Rightarrow \quad F_1 = 0$

F_β–мера: $F_\beta = (1 + \beta^2)\frac{PR}{\beta^2 P + R}$

$P = 1, R = 1 \quad \Rightarrow \quad F_1 = 1$

Нам достаточно $\beta = 1$

$F_1 \in [0,1]$

$$F_1 = 2\frac{PR}{P + R}$$

Существуют и другие меры,
F_1 – одна из множества распространенных

F-мера

Модель логистической регрессии: $0 \leq h(x) \leq 1$

Предсказываем 1, если $h(x) \geq \tau$

Предсказываем 0, если $h(x) < \tau$

При различных значениях τ мы получим разные величины точности P и полноты R.

Как подобрать наилучшее значение для порога τ?

- ☒ Вычислить значения P и R на **тестовом** наборе данных и выбрать значение τ при котором достигается максимум $\frac{P+R}{2}$
- ☒ Вычислить значения P и R на **валидационном** наборе данных и выбрать значение τ при котором достигается максимум $\frac{P+R}{2}$
- ☒ Вычислить значения P и R на **тестовом** наборе данных и выбрать значение τ при котором достигается максимум $2\frac{PR}{P+R}$
- ☑ Вычислить значения P и R на **валидационном** наборе данных и выбрать значение τ при котором достигается максимум $2\frac{PR}{P+R}$

Использование больших данных

«**Большие данные**» (Big Data) – подход в машинном обучении, когда для настройки алгоритма используются очень большие обучающие наборы (миллионы, миллиарды и более) – получил большие распространение в XXI веке.

«Побеждает не лучший алгоритм, а тот, у кого было больше данных»
(из жизни специалистов по ИИ и ML)

Большой объем данных не всегда работает:

Если модель имеет высокое смещение (проблема смещения) →
увеличение объема данных не поможет.

Для BigData необходимо выполнение ряда условий

Исследование Бэнко и Брилл, 2001

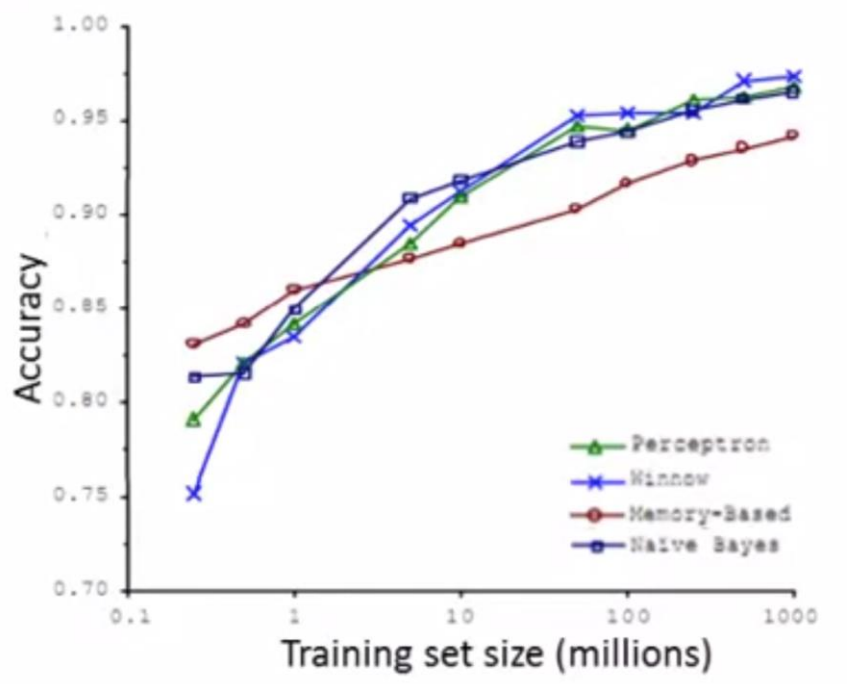
Классификация одинаково звучащих (в речи) слов на английском языке:

two, to, too

For breakfast I ate ____ eggs.

Использовались алгоритмы:

- Персептрон
- Алгоритм просеивания (устаревший)
- На основе памяти (устаревший)
- Наивный байесовский классификатор



Выводы:

1) Все алгоритмы примерно равны
2) Чем больше данных, тем выше достоверность

Особенности:
задача сложна,
большой размер вектора *x*

Использование больших данных

Когда использование BigData оправдано?

Когда вектор параметров *x* содержит достаточно информации для точного предсказания *y*

Интуитивно: Если человек-эксперт на основе этой информации *x* может дать уверенное предсказание *y*, то ее скорее всего достаточно.

Задача Бэнко и Брилл: For breakfast I ate ____ eggs. two, to, too — **Достаточно**

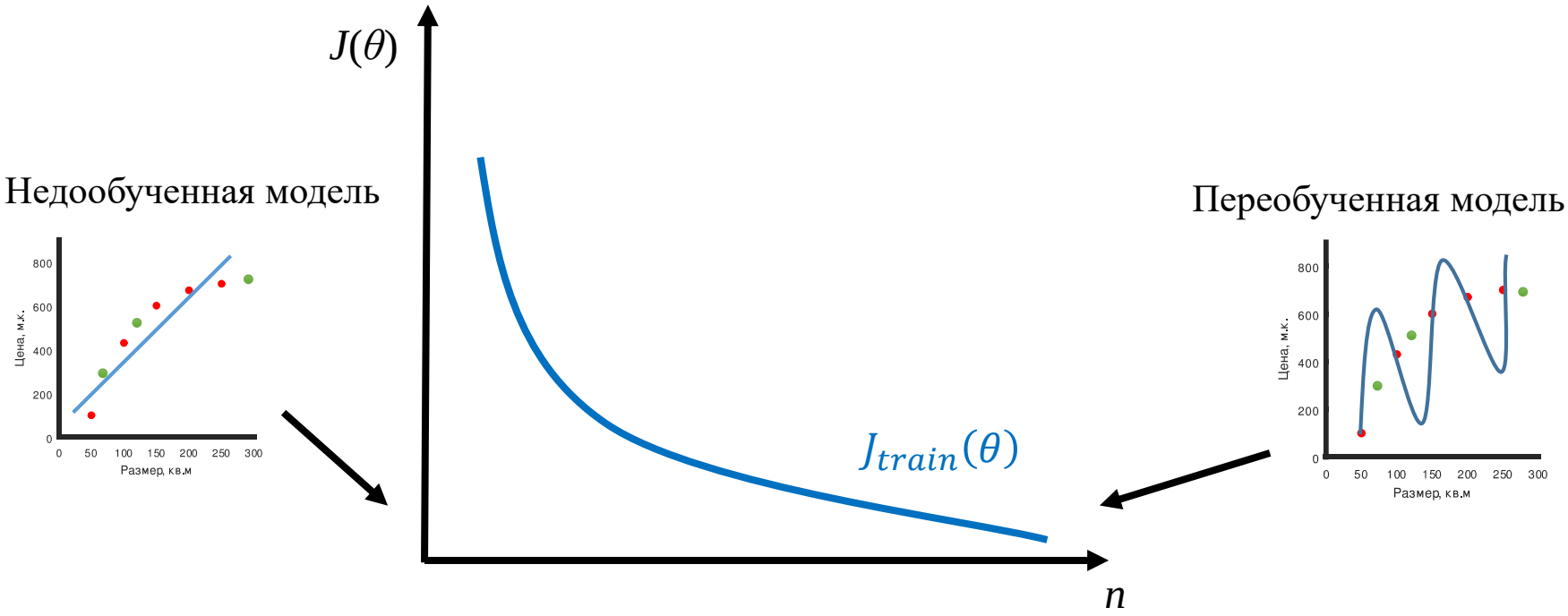
Задача предсказания стоимости дома только по его площади — **Не достаточно**

Использование больших данных

Почему BigData работает?

Для решения сложной задачи используем модель с большим числом параметров (линейная/логистическая регрессия с большим числом признаков; нейронная сеть с множеством скрытых слоев и нейронов)

Модель с малым смещением и возможно большой дисперсией ⇒ $J_{train}(\theta)$ будет мало



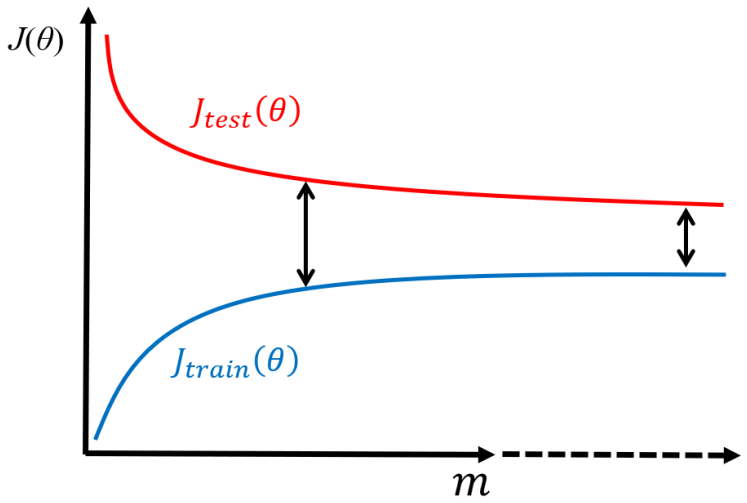
Использование больших данных

Почему BigData работает?

Для решения сложной задачи используем модель с большим числом параметров (линейная/логистическая регрессия с большим числом признаков; нейронная сеть с множеством скрытых слоев и нейронов)

Модель с малым смещением и возможно большой дисперсией → $J_{train}(\theta)$ будет мало

Для устранения проблемы дисперсии один из хороших путей – увеличение объема данных (уменьшает вариабельность)



$J_{test}(\theta) \approx J_{train}(\theta)$

↓

$J_{test}(\theta)$ будет мало

Использование больших данных

Какие задачи не решить с помощью BigData?

Задачи с большим смещением

- Задача предсказания стоимости дома только по его площади

  Стоимость зависит еще от большого числа факторов (возраст, число комнат, этаж и пр.)

- Задача предсказания курса акций/валюты на основе прошлых значений

  Стоимость зависит от политической, экономической, социальной ситуации, но не от старых значений

- Задача распознавания эмоций человека по фотографии

  Статического изображения не достаточно для полноценной идентификации эмоций

Если исходных данных для решения **не достаточно** (вектор параметров *x* не содержит достаточно информации для точного предсказания *y*, человек эксперт не может дать верное решение по этой информации), то задача очень вероятно **не решается** с привлечением больших данных.

Использование больших данных

В каких ситуациях из перечисленных ниже использование больших данных вероятно **не поможет** решить задачу?

- ☑ Признаки *x* не содержат достаточно информации для достоверного предсказания
- ☒ Используется алгоритм с большим числом параметров
- ☑ Признаки *x* не содержат достаточно информации для достоверного предсказания, но используется алгоритм с большим числом параметров
- ☒ Не используется регуляризация (или $\lambda = 0$)