

# Математический анализ данных и машинное обучение

Лекция 1

Саркисян Вероника

# Коммуникация

Все материалы (презентации лекций, ноутбуки с семинаров, данные, домашние задания) будут выкладываться на [GitHub](#).

Все вопросы можно задавать по почте ([impecopeco@gmail.com](mailto:impecopeco@gmail.com))

In case of emergency: 8-926-827-35-45

# Структура курса

Модуль 1	Классическая постановка задач машинного обучения. Обзор алгоритмов и задач ML. Задача линейной регрессии. Метрики качества в задаче регрессии.
Модуль 2	Задача классификации. Логистическая регрессия. Метрические классификаторы: KNN, SVM. Метрики качества в задаче классификации.
Модуль 3	Решающие деревья. Ансамбли и композиции. Бутстреп и бэггинг. XGBoost, CatBoost.
Модуль 4	Обучение без учителя. Методы понижения размерности признакового пространства и визуализация. Препроцессинг данных.
Модуль 5	Защита проектов.

# Оценки за курс

## Формы контроля:

1. Еженедельные домашние задания по практическому семинарскому материалу
2. Проверочные работы по материалу лекций (зачет/незачет)
3. Финальный проект

## Итоговая оценка:

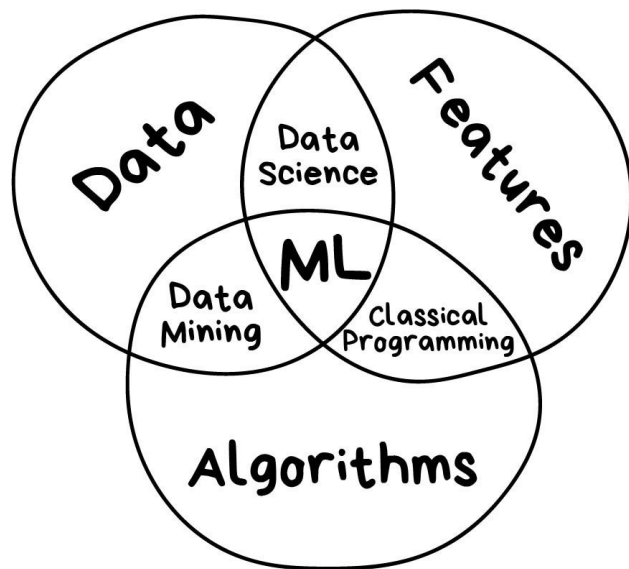
**$\text{ДЗ} * 0.5 + \text{Тесты} * 0.2 + \text{Проект} * 0.3$**

Если все проверочные работы сданы на зачет, и по всем ДЗ набрано не меньше 0.8 от максимального балла, слушатель может зачесть накопленную оценку  $((\text{ДЗ} * 0.5 + \text{Тесты} * 0.2)/0.7)$  автоматом, без сдачи проекта.

# План на сегодня

13:30 - 15:00	Вводная лекция: что такое машинное обучение? Тест.
15:15 - 16:00	Лекция: линейная регрессия и метрики качества в задаче регрессии.
16:10 - 16:40	Перерыв
17:00 - 18:00	Семинар: знакомимся с основными инструментами.
18:00 - 20:00	Семинар: решаем задачи, знакомимся с Kaggle.

# Что такое машинное обучение?



# Постановка задачи машинного обучения

$X$  - множество объектов; каждый объект  $x \in X$  представлен вектором признаков  $(f_1, f_2, \dots, f_n)$

$Y$  - множество допустимых ответов

$L(y, \hat{y})$  - функция ошибки (аргументы - истинные ответы и оцененные); по умолчанию чем больше значение функции, тем больше ошибка алгоритма.

Задача состоит в том, чтобы построить алгоритм  $a: X \rightarrow Y$ , минимизирующий функцию ошибки  $L$ .

# Матрица объекты-признаки

X

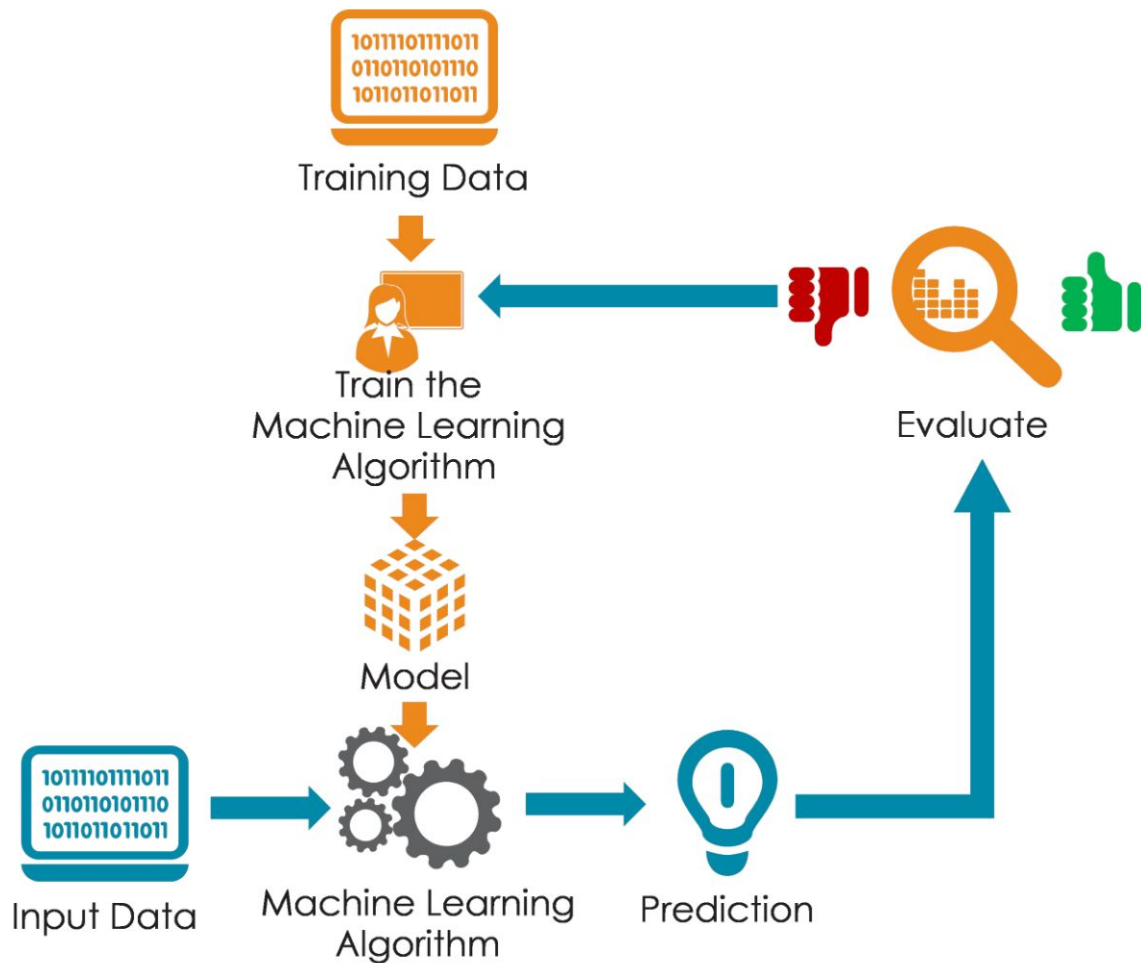
$y^*$

features

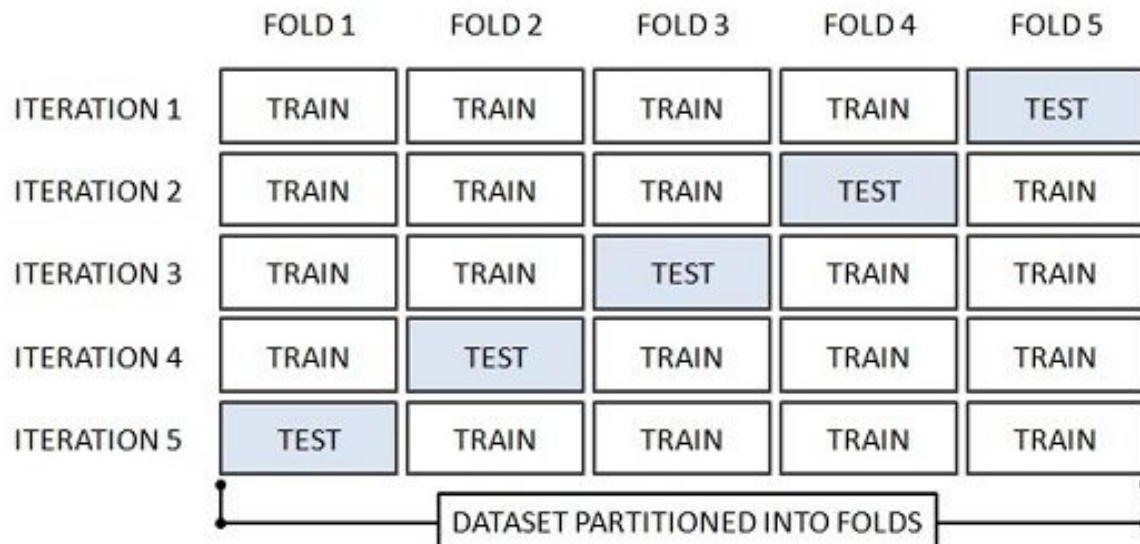
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
6	0	3	Moran, Mr. James	male		0	0	330877
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

$Y=\{0,1\}$



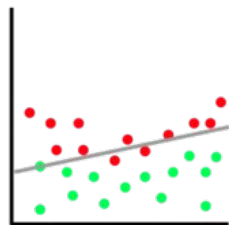


# Отложенная выборка и кросс-валидация

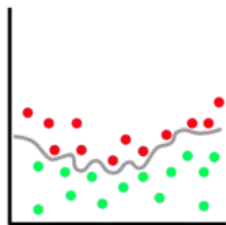


# Переобучение

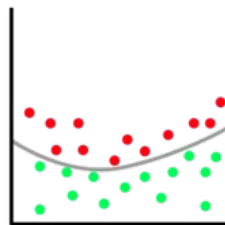
Переобучение - это потеря способности модели к обобщению: качество модели на тренировочной выборке высокое, а на тестовой - значительно ниже.



Underfitting



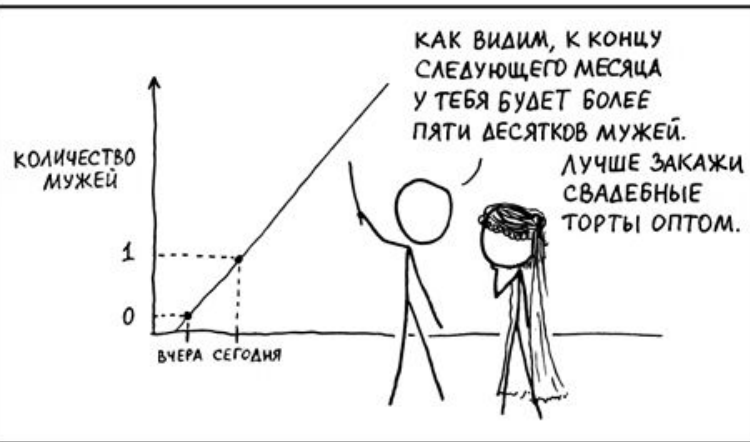
Overfitting



Balanced

learning & regularization

МОЁ ХОББИ: ЭКСТРАПОЛИРОВАТЬ



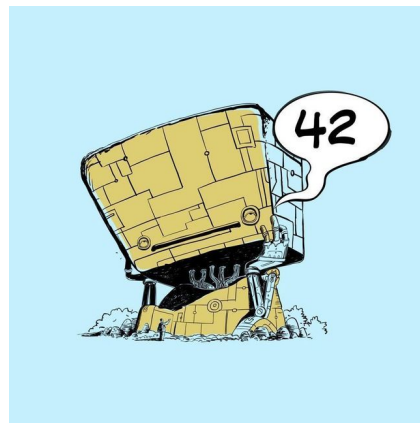
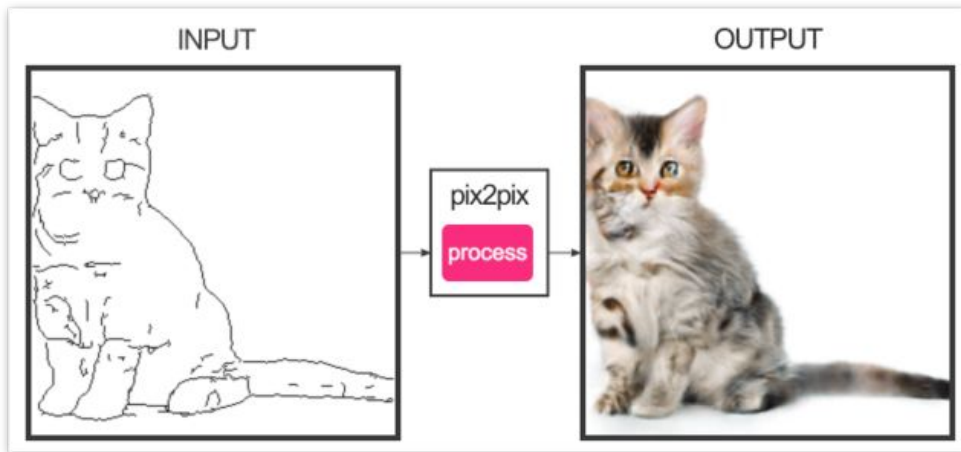
# Что “умеют” модели машинного обучения?

## Умеют:

- Предсказывать некоторую переменную
- Извлекать зависимости из данных
- Обобщать
- Генерировать объекты, аналогичные данному

## Не умеют:

- Создавать качественно новые объекты
- Решать новый, неизвестный модели класс задач
- Отвечать на главный вопрос жизни, вселенной и всего такого





“Зоопарк” моделей  
машинного обучения

# Классическое Обучение



Почта НОВЫХ НЕТ

Написать письмо

Диск

[Получить Плюс](#)

Новомосковский



На Хабарова при въезде в  
Град Московский строят  
какой то домик и парковку  
рядом с ним. А та же вид...

Konstantin

7

SOYUZ.RU



В сиквеле хоррора «Оно»  
будет один из самых  
странных моментов книги  
Стивена Кинга

## Сейчас в СМИ в Москве 16 сентября, воскресенье 19:57

СМИ сообщили о двух новых подозреваемых в деле об отравлении Скрипалей

Кандидат от КПРФ стал лидером во втором туре выборов в Приморье

Украина создаст военную базу на Азовском море до конца года

В Петербурге прошел митинг против пенсионной реформы

Ротенберг уверен, что Крымский мост простоят сто лет

USD MOEX 68,05 -0,20 EUR MOEX 79,11 -0,67 НЕФТЬ 78,11 -0,26% ...

## Детские товары

Проверенные скидки до 50% на Маркете

Яндекс

[Видео](#) [Картинки](#) [Новости](#) [Карты](#) [Маркет](#) [Переводчик](#) [Музыка](#) [ТВ онлайн](#) [ещё](#)

Найдётся всё. Например, когда цветет амброзия



Установите Яндекс.Браузер

ДОМА И В ПОЕЗДКАХ ПО РОССИИ:  
БЕЗЛИМИТНЫЙ ИНТЕРНЕТ И ЗВОНОК НА ВСЕХ ОПЕРАТОРОВ

ПОСМОТРЕТЬ 5М6

## Погода

 +14° Ночью +8,  
утром +11

## Пробки

1 До конца дня дороги свободны

## Карта Москвы

[Метро](#) [Такси](#) [Расписания](#)  
[Сообщество соседей](#)

## Посещаемое

[Маркет](#) — духи и туалетная вода[Авто.ру](#) — пятилетние иномарки[Недвижимость](#) — квартиры на 1 этаже[ТВ онлайн](#) — смотреть МУЗ-ТВ[Деньги](#) — оплата света и волы

## Телепрограмма

[▶ ТВ онлайн](#)19:25 [Лучше всех!](#) [Первый](#)19:27 [Как извести любовницу...](#) [ТВ Центр](#)19:30 [Спец](#) [Пятый канал](#)19:30 [Новости культуры](#) [С...](#) [Культура](#)

## Афиша

[Хищник](#) [премьера](#)[История одного назначения](#) [драма](#)[Гоголь. Страшная месть](#) [детектив](#)[Великий уравниль](#) [2 боевик](#)

Сейчас в СМИ в Москве 16 сентября, воскресенье 19:57

СМИ сообщили о двух новых подозреваемых в деле об отравлении Скрипалей

Кандидат от КПРФ стал лидером во втором туре выборов в Приморье

Украина создаст военную базу на Азовском море до конца года

В Петербурге прошел митинг против пенсионной реформы

Ротенберг уверен, что Крымский мост простоят сто лет

USD MOEX 68,05 -0,20 EUR MOEX 79,11 -0,67 НЕФТЬ 78,11 -0,26% ...



Детские товары

Проверенные скидки до 50% на Маркете

Видео Картинки Новости Карты Маркет Переводчик Музыка ТВ онлайн ещё

Яндекс

ранжирование результатов

поиска пример, когда цветет амброзия



Найти



Установите Яндекс.Браузер



ТАРИФИЩЕ

ДОМА И В ПОЕЗДКАХ ПО РОССИИ:  
БЕЗЛИМИТНЫЙ ИНТЕРНЕТ И ЗВОНКИ НА ВСЕХ ОПЕРАТОРОВ

ПОСМОТРЕТЬ 5М6

Погода

+14° Ночью +8, утром +11

Посещаемое

Маркет — духи и туалетная вода

Авто.ру — пятилетние иномарки

Недвижимость — квартиры на 1 этаже

ТВ онлайн — смотреть МУЗ-ТВ

Деньги — оплата света и воды

Пробки

1 До конца дня дороги свободны

Телепро

19:25 Лучший

19:27 Как известилась... ТВ Центр

19:30 Спец. Пятый канал

19:30 Новости культуры С... Культура

Карта Москвы

Метро Такси Расписания

Сообщество соседей

Афиша

Хищник премьера

История одного назначения драма

Гоголь. Страшная месть детектив

Великий уравнильщик боевик

регрессия:  
предсказываем  
погоду

регрессия:  
предсказываем  
пробки

кластеризация:  
определяем, к  
какой группе по  
интересам  
относится  
пользователь.



Настройка

veronica.fkn



Получить Плюс

Получить Плюс

московский

На Хабарова при въезде в  
Град Московский строят  
какой то домик и парковку  
рядом с ним. А та же вид...

Konstantin

7

SOYUZ.RU



В сиквеле хоррора «Оно»  
будет один из самых  
странных моментов книги  
Стивена Кинга



Тест

**Соотнесите данные задачи с задачами машинного обучения: назовите тип задачи, что является объектами и целевой переменной.**

1. Определение возраста человека по фотографии
2. Определение пола человека по фотографии
3. Кредитный скоринг (оценка кредитоспособности клиента)
4. Распознавание рукописного текста (скан -> txt)
5. Определение тональности текста (позитивная/негативная/нейтральная)
6. Разделение пользователей форума на группы по интересам
7. Фильтрация спама
8. Определение жанра фильма

Номер задачи	Тип задачи	Объект	Целевая переменная
0: предсказание зарплаты человека по его резюме	Регрессия	Резюме	Зарплата

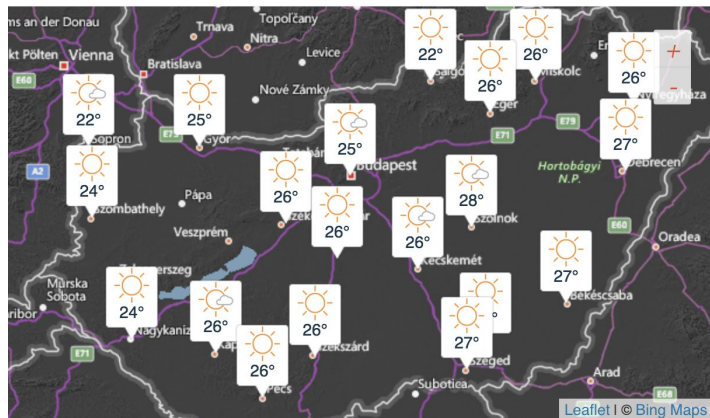
Перерыв

# Задача регрессии

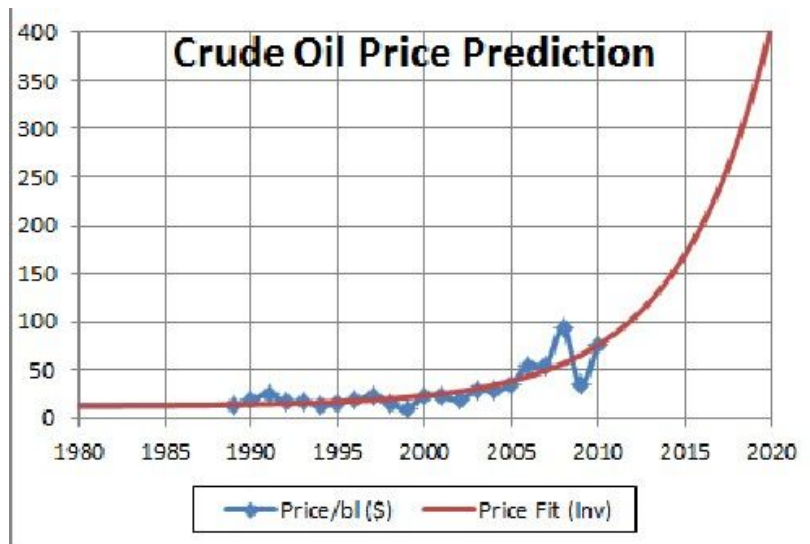
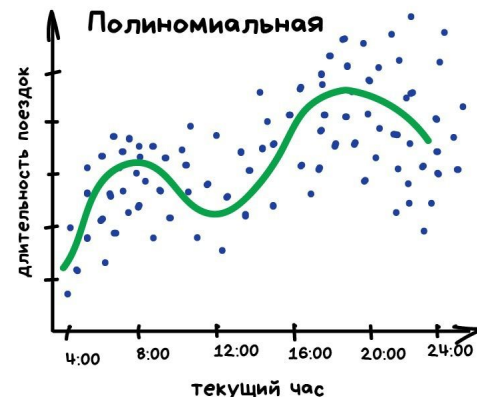
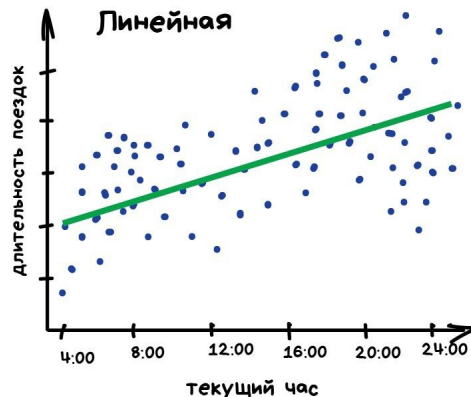
Предсказываем числовую переменную.

- Линейная регрессия
- Полиномиальная регрессия

Венгрия: карта погоды



Предсказываем пробки



Регрессия

# Постановка задачи линейной регрессии

Алгоритм:

$$a(x) = w_0 + \sum_{j=1}^d w_j x_j. \quad \Rightarrow \quad a(x) = w_0 + \langle w, x \rangle, \quad \Rightarrow \quad a(x) = \langle w, x \rangle.$$

Метрики качества:

mean squared error

root mean squared error

mean absolute error

$$\text{MSE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2. \quad \text{RMSE}(a, X) = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2}. \quad \text{MAE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|.$$

# Решение задачи регрессии

Зафиксируем функционал ошибки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w$$

Перепишем задачу в матричном виде:

$$\frac{1}{\ell} \|Xw - y\|^2 \rightarrow \min_w,$$

Решение в явном виде:

$$w = (X^T X)^{-1} X^T y.$$

# Регуляризация

К функционалу качества  $Q$  добавим слагаемое, штрафующее за большие веса:

$$Q_\alpha(w) = Q(w) + \alpha R(w).$$

Наиболее распространенными являются  $L_2$  и  $L_1$ -регуляризаторы:

$$R(w) = \|w\|_2^2 = \sum_{i=1}^d w_i^2,$$

$$R(w) = \|w\|_1 = \sum_{i=1}^d |w_i|.$$

# Инструменты



# GitHub

[Хороший пошаговый tutorial](#) о том как использовать Git из командной строки.

Combo-Breaker / ML\_course

Watch 0Star 0Fork 0

<> Code

Issues 0

Pull requests 0

Projects 0

Wiki

Insights

Settings

Курс МАДМО в Сбербанке.

Edit

[Manage topics](#)

8 commits

1 branch

0 releases

1 contributor

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download

Combo-Breaker Wine Dataset

Latest commit df44a58 19 minutes ago

Lecture

Create План

23 minutes ago

Seminar

Wine Dataset

19 minutes ago

README.md

Update README.md

3 days ago

README.md

ML\_course

Математический анализ данных и машинное обучение.

# Python

NumPy - библиотека для удобной работы с векторами, матрицами и т.д.

Pandas - библиотека для работы с датасетами.


Matplotlib - библиотека для визуализации данных.

Sklearn - библиотека с реализацией основных алгоритмов машинного обучения.

[Anaconda](#)

# Jupyter Notebook

## Jupyter Notebook

Jupyter English\_bilstm\_NER Last Checkpoint: 14.12.2017 (autosaved)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Code

```
In [1]: import pandas as pd

df = pd.read_csv('ner.csv', encoding = "ISO-8859-1", error_bad_lines=False)

df.head()
```

b'Skipping line 281837: expected 25 fields, saw 34\n'

Out[1]:

	Unnamed: 0	lemma	next-lemma	next-next-lemma	next-next-pos	next-next-shape	next-next-word	next-pos	next-shape	next-word	...	prev-prev-lemma	prev-prev-pos	prev-prev-shape	prev-prev-word
0	0	thousand	of	demonstr	NNS	lowercase	demonstrators	IN	lowercase	of	...	__start2__	__START2__	wildcard	__START
1	1	of	demonstr	have	VBP	lowercase	have	NNS	lowercase	demonstrators	...	__start1__	__START1__	wildcard	__START
2	2	demonstr	have	march	VPN	lowercase	marched	VBP	lowercase	have	...	thousand	NNS	capitalized	Thousai
3	3	have	march	through	IN	lowercase	through	VPN	lowercase	marched	...	of	IN	lowercase	
4	4	march	through	london	NNP	capitalized	London	IN	lowercase	through	...	demonstr	NNS	lowercase	demonstrat

5 rows x 25 columns