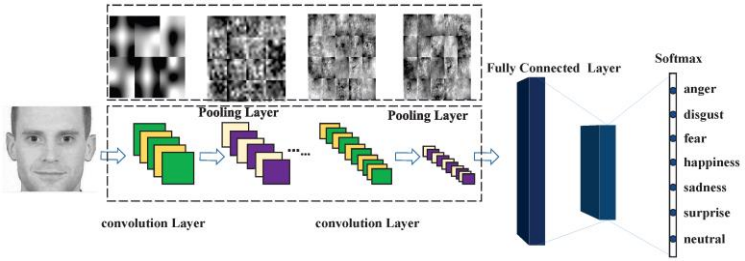
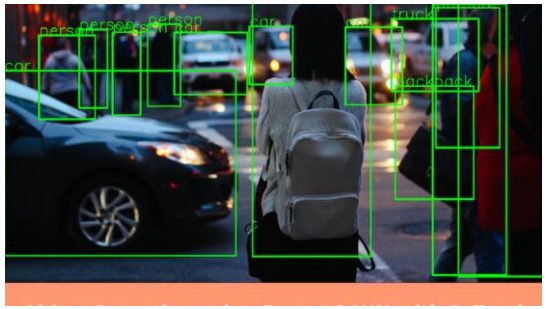
Основы искусственного интеллекта





Лекция 2

Процедура обучения моделей на основе машинного обучения. Метрики качества

к.ф.-м.н., доцент кафедры ИСиЦТ
Корнаева Е.П.

Процедура обучения моделей на основе машинного обучения

Целевая функция – функция, подлежащая минимизации (максимизации)

В ML целевая функция называется функцией потерь (loss function)

Параметры модели определяются в ходе решения задачи МО.
Например, в задачах регрессии параметрами являются компоненты матрицы весовых коэффициентов $\boldsymbol{\Theta}$.
Гиперпараметры задаются пользователем, как правило не единственным образом, и их значения влияют на значения искомых параметров.

1. Масштабирование признаков / Feature Scaling
2. Скорость обучения α / Learning rate
3. Погрешность δ и количество итераций $N_{\max}$ / Error and # of iterations
4. Регуляризация / Regularization

Процедура обучения моделей на основе машинного обучения

Общая процедура построения приближенных моделей

Процедура валидации:

На отложенной выборке

Подбор гиперпараметров:

- Построить *l*-моделей для каждого значения гиперпараметра по обучающей выборке;
- Рассчитать точность моделей на новой выборке (проверочной, валидационной);
- Выбрать значение *гиперпараметра*, для которого точность максимальная.

Кросс-валидация

fold = 3

Каждая из *l* моделей строиться *fold* раз. Считается средняя точность по всем *fold* моделям.
Выбирается модель с максимальной точностью.
Обычно fold (блок) = 3 (5)

Процедура тестирования:
Проверяется точность выбранной модели на новой (тестовой выборке)!!!

Процедура обучения моделей на основе машинного обучения

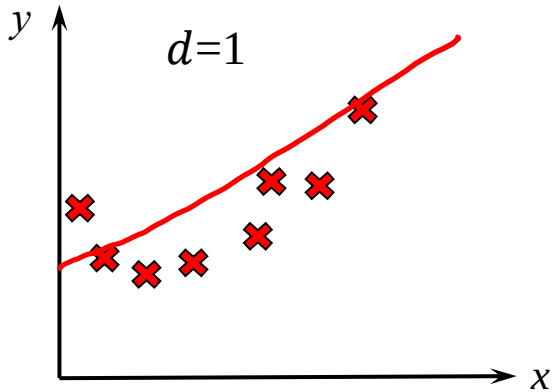
Параметры модели определяются в ходе решения задачи МО. Например, в задачах регрессии параметрами являются компоненты матрицы весовых коэффициентов $\boldsymbol{\theta}$.
Гиперпараметры задаются пользователем, как правило не единственным образом, и их значения влияют на значения искомых параметров.

1. Масштабирование признаков / Feature Scaling
2. Скорость обучения α / Learning rate
3. Погрешность δ и количество итераций $N_{\max}$ / Error and # of iterations
4. Регуляризация / Regularization

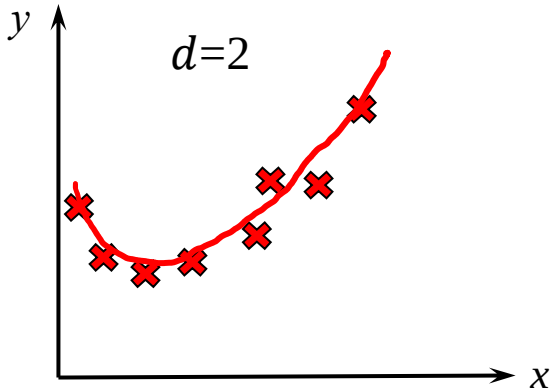
$$J(\theta_0, \theta_j) = \frac{1}{2n}\left[\sum_{i=1}^{n}(H(\theta, X_i) - Y_i)^2 + \lambda \sum_{j=1}^{k}\theta_j^2\right] \Rightarrow \min$$

y_{train}
y_{test}

$h(\theta_0, \theta_j) = \theta_0 + \theta_p x^p,$
$(j, p = 1 \ldots d).$

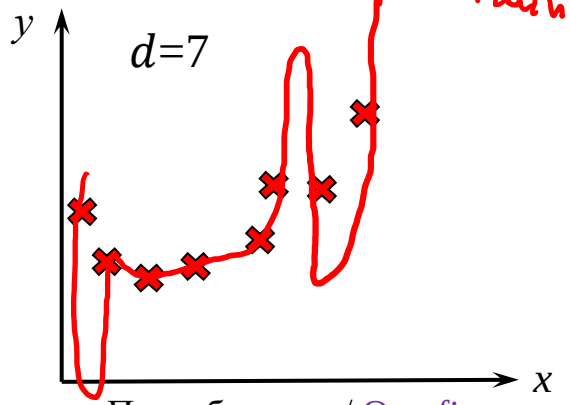


Недообучение / Underfit, High bias



Норм. / Appropriate fitting

$y_{train} = 0$



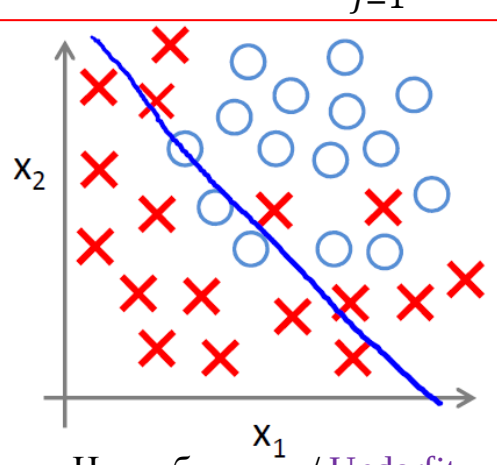
Переобучение / Overfit, High variance

Процедура обучения моделей на основе машинного обучения

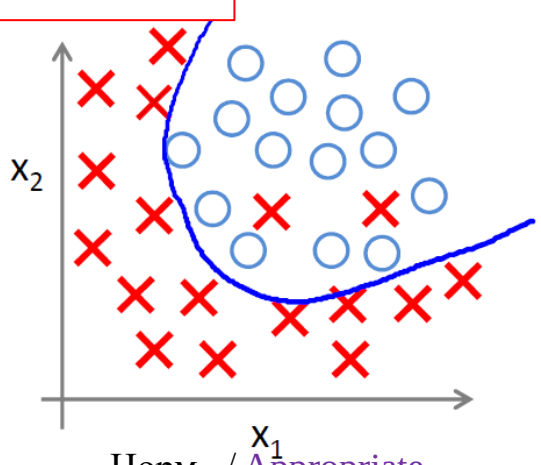
Параметры модели определяются в ходе решения задачи МО. Например, в задачах регрессии параметрами являются весовые коэффициенты Θ.
Гиперпараметры задаются пользователем, как правило не единственным образом, и их значения влияют на значения искомых параметров.

1. Масштабирование признаков / Feature Scaling
2. Скорость обучения α / Learning rate
3. Погрешность δ и количество итераций $N_{\max}$ / Error and # of iterations
4. Регуляризация / Regularization

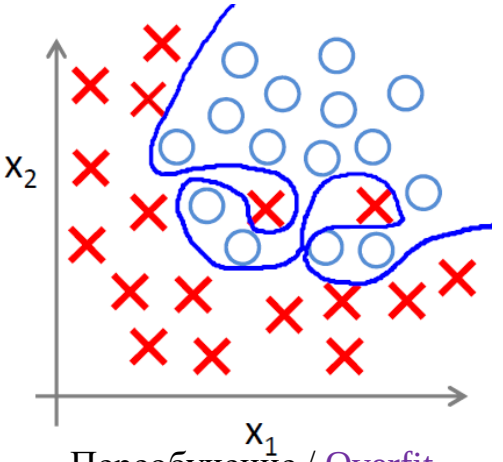
$$J(\Theta) = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \ln(H(\theta, X_i)) + (1 - Y_i)(\ln(1 - H(\theta, X_i))) + \frac{\lambda}{2n}\sum_{j=1}^{k}\theta_j^2 \Rightarrow \min$$



Недообучение / Underfit, High bias



Норм. / Appropriate fitting



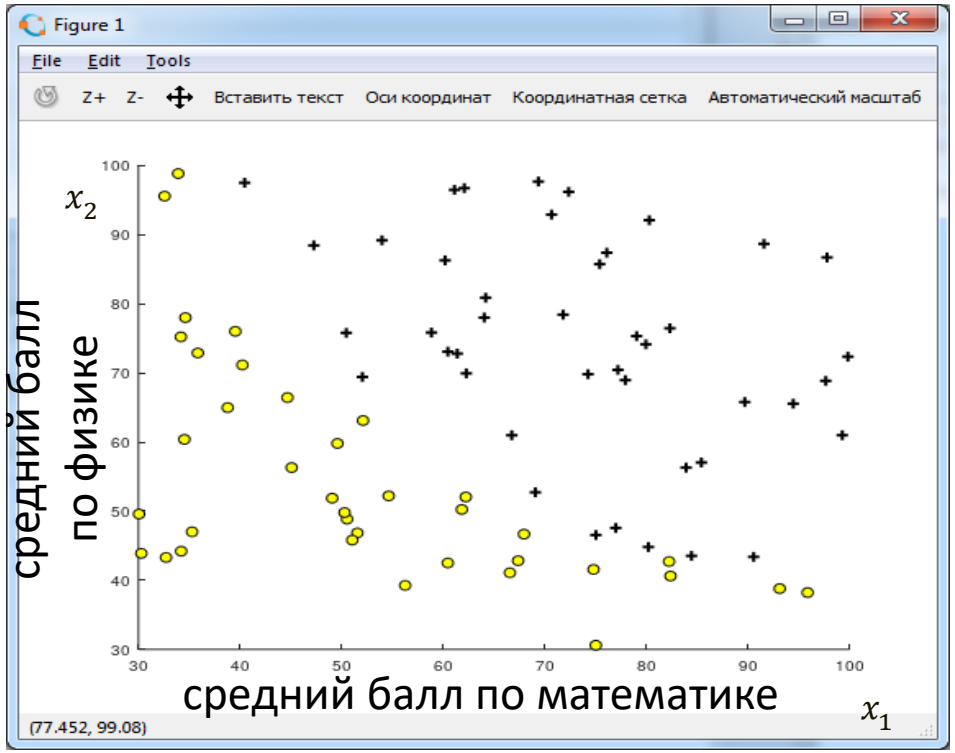
Переобучение / Overfit, High variance

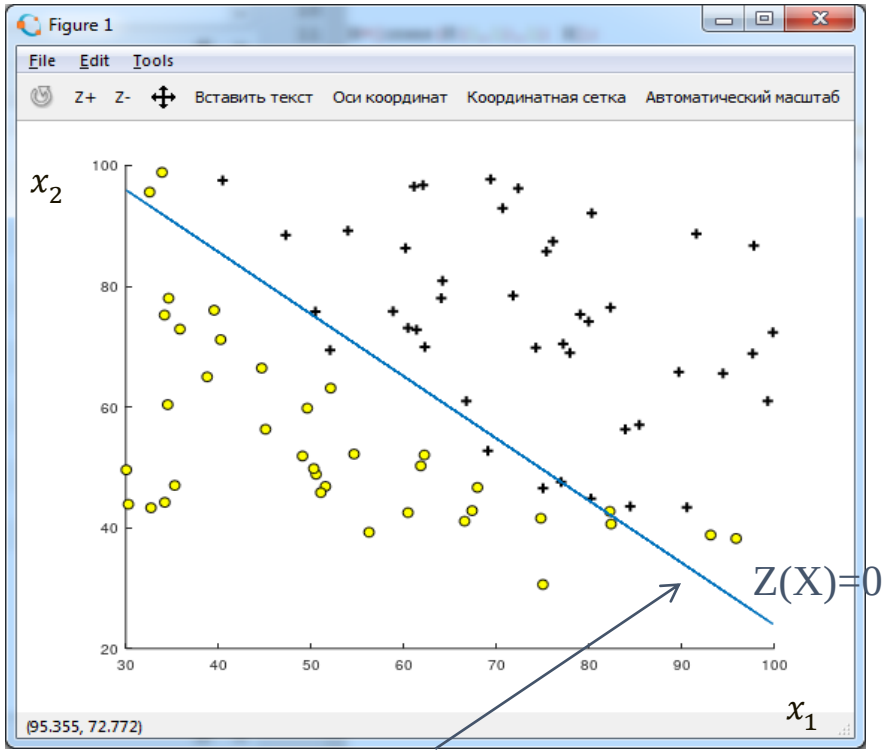
Процедура обучения моделей на основе машинного обучения

Метод k-ближайших соседей (KNN) для задачи классификации

Пример бинарной классификации: $Y \in \{0; 1\}$

Двумерный случай: $\boldsymbol{X} = [[x_1, x_2]]$





+ $y_i = 1$: зачислен ○ $y_i = 0$: не зачислен

$z(\boldsymbol{X}) = 0,$

Процедура обучения моделей на основе машинного обучения

Метод k-ближайших соседей (KNN) для задачи классификации

Обучающая выборка: $\{(X_{ij}, Y_i)\}$

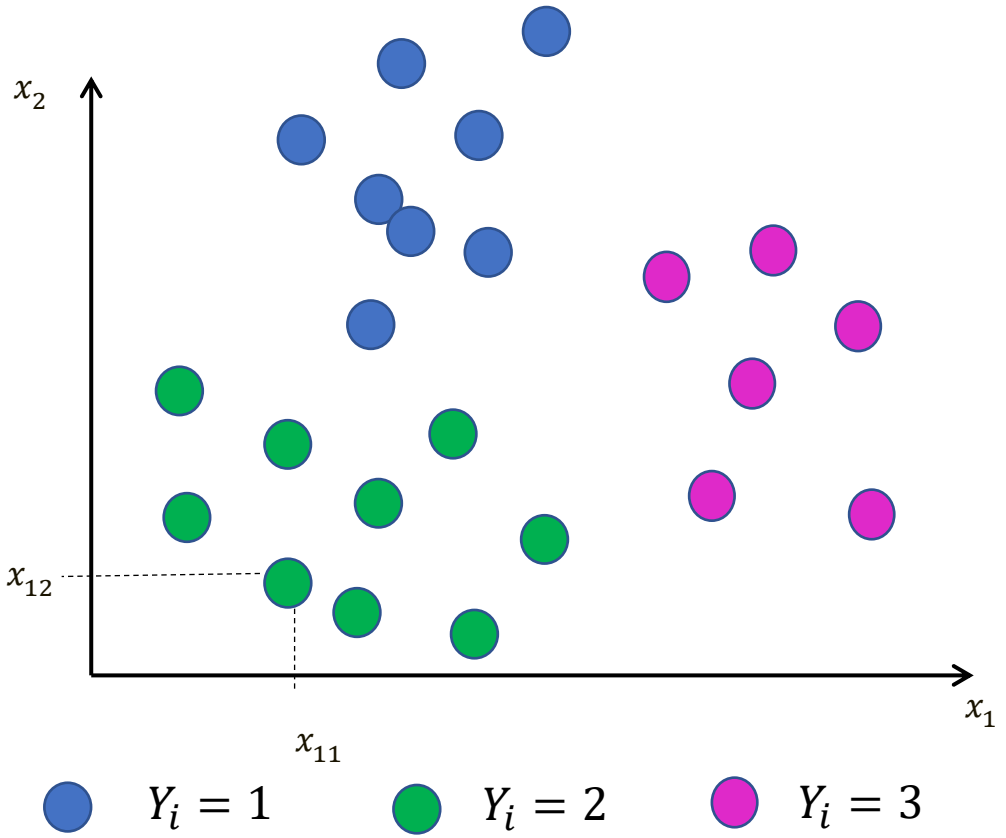
i – номер объекта; j – номер признака;
X_{ij} – значение $j^{\text{го}}$ признака для $i^{\text{го}}$ объекта;
Y_i – значение класса для $i^{\text{го}}$ объекта (дискретная величина);
n – кол-во объектов;
m – кол-во признаков (факторов);

Матричная форма записи:

	X_1	X_2	…	X_m	$\boldsymbol{Y}_{[n\times 1]}$
1	x_{11}	x_{12}		x_{1m}	y_1
2	x_{21}	x_{22}		x_{2m}	y_2
…			…		
n	x_{n1}	x_{n2}		x_{nm}	y_n

$\boldsymbol{X}_{[n\times(m+1)]}$

Пример

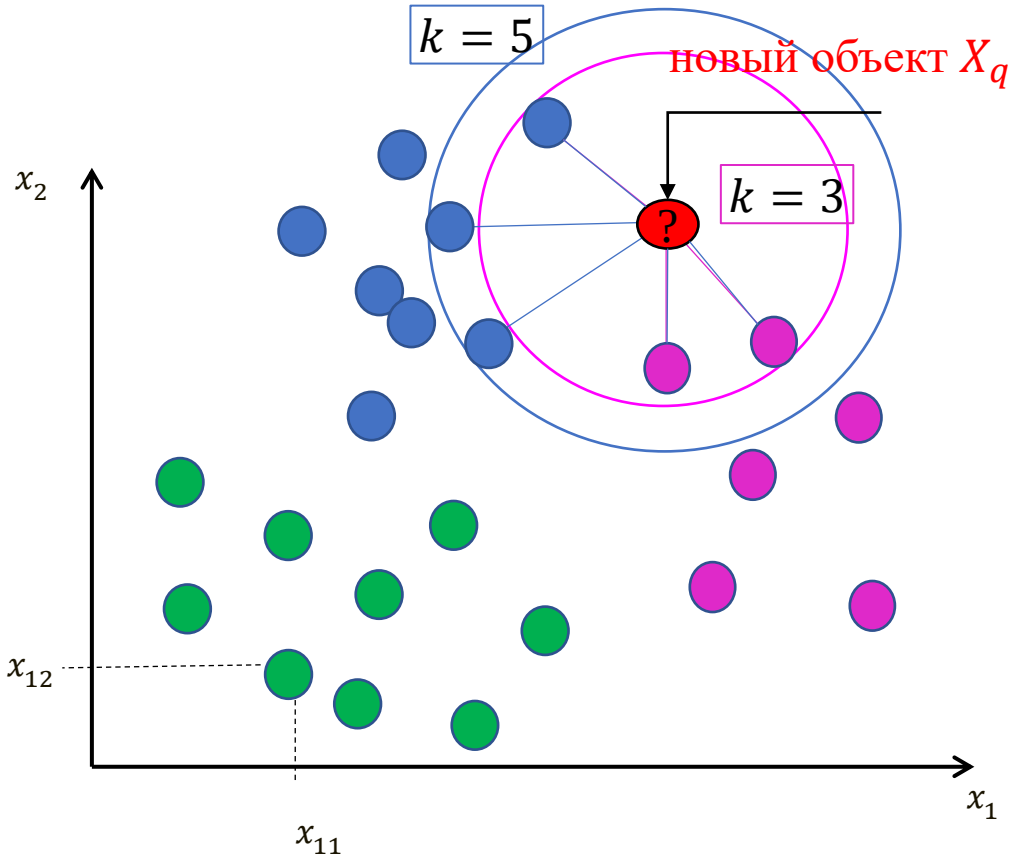


	X_1	X_2	Y
1	x_{11}	x_{12}	y_1
2	x_{21}	x_{22}	y_2
…			
n	x_{n1}	x_{n2}	y_n

Процедура обучения моделей на основе машинного обучения

Метод k-ближайших соседей (KNN) для задачи классификации

✓ Гипотеза компактности: предположение о том, что схожие объекты гораздо чаще лежат в одном классе, чем в разных



Обучение:

- Сохраняется обучающая выборка $\{X_i, Y_i\}$;

Классификация нового объекта:

- Измерить расстояние от всех объектов до нового объекта X_q;
- Упорядочить объекты в порядке возрастания (неубывания) дальности до нового объекта:

$$\rho(X_l, X_q) \leq \cdots \leq \rho(X_i, X_q) \leq \ldots \leq \rho(X_r, X_q)$$

- Выбрать первые k объектов (k ближайших соседей):

$$\{X_1, \ldots, X_k\}$$

- Назначить новому объекту модальный класс (самый частый) среди k ближайших соседей:

$$Y(X_q) = \underset{y_{cl} \in Y}{\operatorname{argmax}} \sum_{i=1}^{k} [y_i == y_{cl}]$$

Процедура обучения моделей на основе машинного обучения

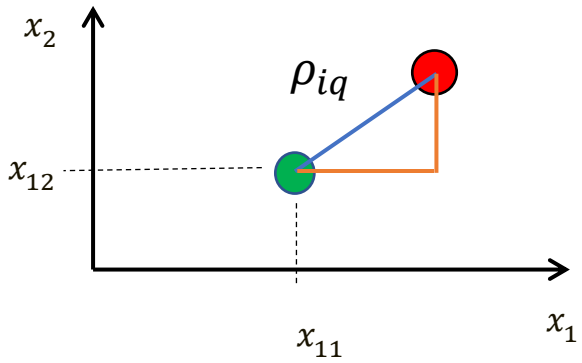
Метод k-ближайших соседей (KNN) для задачи классификации

Метрика (в функциональном анализе) - это функция с двумя аргументами, принимающая значения в множестве неотрицательных вещественных чисел, удовлетворяющая условиям:

$\rho(x,z) = 0 \leftrightarrow x = z$ – аксиома тождества

$\rho(x,z) = \rho(z,x)$ - аксиома симметрии

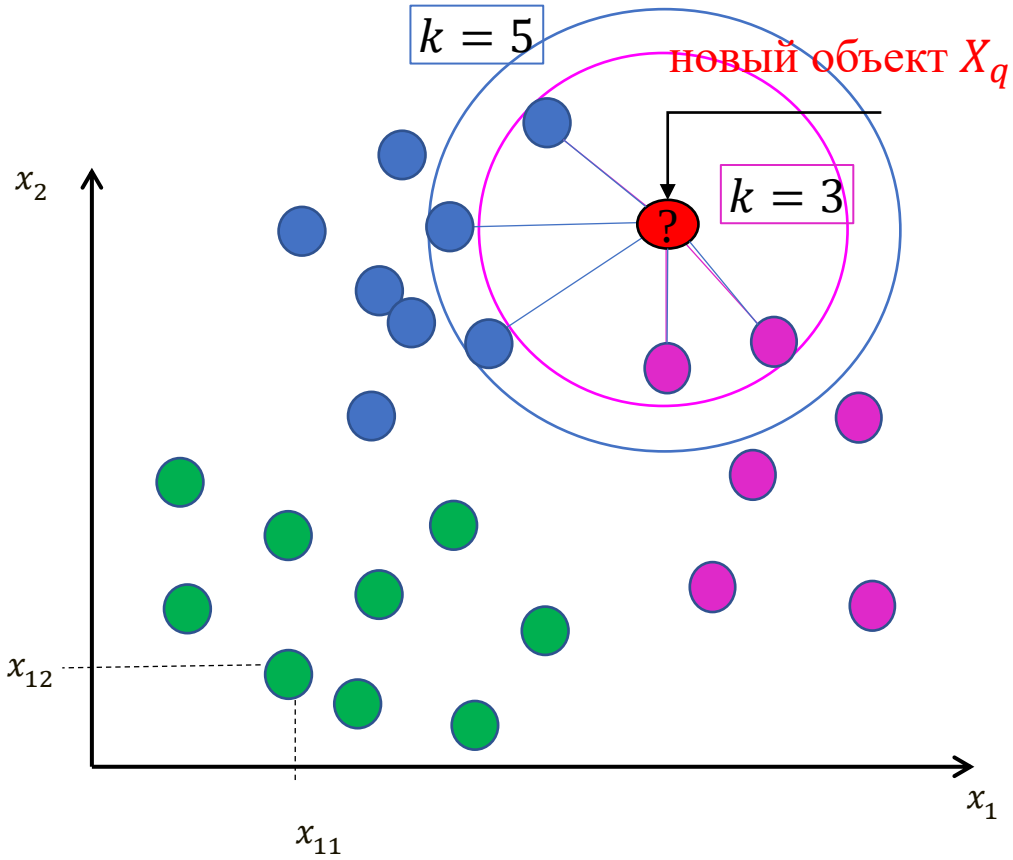
$\rho(x,z) \leq \rho(x,v) + \rho(v,z)$ - неравенство треугольника



Вид расстояний	Расчет $\rho_{iq} = \rho(X_i, X_q)$	Примечание
Евклидова норма	$\sqrt{\sum_{j=1}^{m}(x_{ij} - x_{iq})^2}$	Самое частое применение
Степенное расстояние	$\left(\sum_{j=1}^{m}(x_{ij} - x_{iq})^p\right)^{1/p}$	Параметр p подбираются в процессе решения.
Расстояние Чебышева	$\max_j \lvert x_{ij} - x_{iq}\rvert$	объекты различаются по какой-либо одной координате (признаку)
L1 – метрика (Манхэттенское расстояние)	$\sum_{j=1}^{m} \lvert x_{ij} - x_{iq}\rvert$	влияние отдельных больших разностей (выбросов) уменьшается

Процедура обучения моделей на основе машинного обучения

Метод k-ближайших соседей с весами (KNN) для задачи классификации



Обучение:

- Сохраняется обучающая выборка $\{X_i, Y_i\}$;

Классификация нового объекта:

- Измерить расстояние от всех объектов до нового объекта X_q;
- Упорядочить объекты в порядке возрастания дальности до нового объекта:

$$\rho(X_l, X_q) \leq \cdots \leq \rho(X_i, X_q) \leq \ldots \leq \rho(X_r, X_q)$$

- Выбрать первые k объектов (k ближайших соседей):

$$\{X_1, \ldots, X_k\}$$

- Назначить новому объекту модальный класс (самый частый) среди k ближайших соседей:

$$Y(X_q) = \underset{y_{cl} \in Y}{\operatorname{argmax}} \sum_{i=1}^{k} \omega_i [y_i == y_{cl}]$$

$w_i = K\left(\frac{\rho(X_i, X_{cl})}{h}\right)$ - Парзеновское окно

K — ядро;
h — ширина окна.

Процедура обучения моделей на основе машинного обучения

Определение гиперпараметров на примере k-ближайших соседей

k – гиперпараметр модели

Процедура валидации:

На отложенной выборке

Подбор гиперпараметра k:

- Построить *l*-моделей для каждого значения гиперпараметра *k* по обучающей выборке;
- Рассчитать точность моделей на новой выборке (проверочной, валидационной);
- Выбрать значение *k*, для которого точность максимальная.

Кросс-валидация

fold = 3

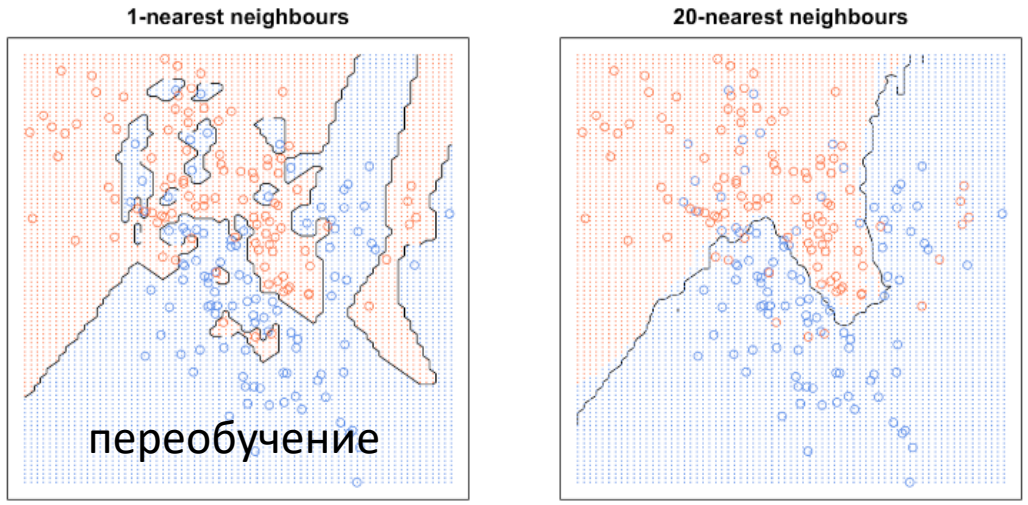
Каждая из *l* моделей строиться *fold* раз. Считается средняя точность по всем *fold* моделям.
Выбирается модель с максимальной точностью.
Обычно кол-во fold (блоков): 3, 5

Процедура тестирования:
Проверяется точность выбранной модели на новой (тестовой выборке)!!!

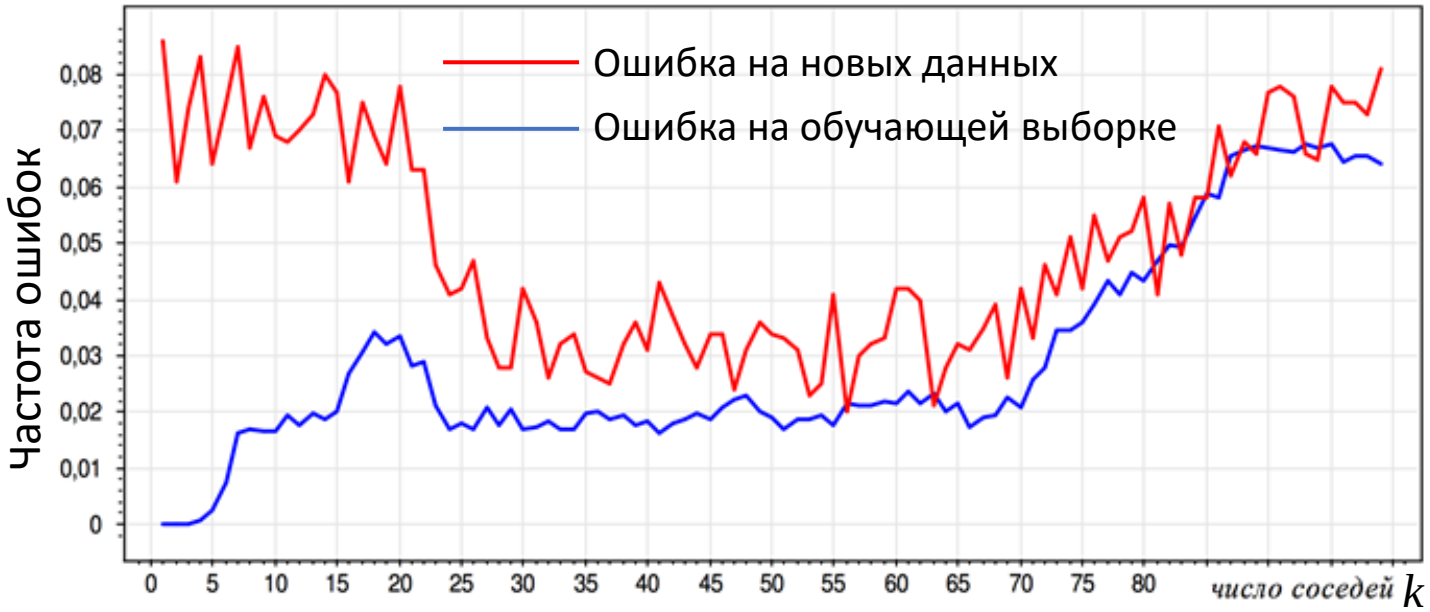
Процедура обучения моделей на основе машинного обучения

Определение гиперпараметров на примере k-ближайших соседей

k – гиперпараметр модели



https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/



Лекции К.В. Воронцова
http://www.machinelearning.ru/wiki/index.php?title=MO

Процедура обучения моделей на основе машинного обучения

Общая процедура построения приближенных моделей

Исходная выборка делится на три части:

- Обучающая (определение параметров модели);
- Валидационная (подбор гиперпараметров на основе валидации или кросс-валидации);
- Тестовая (расчет точности модели).

Процедура обучения моделей на основе машинного обучения

Метрики качества модели классификации

Метрика качества модели (алгоритма обучения) (в машинном обучении) – количественная оценка качества модели (т.е. ее обобщающей способности)

Accuracy – относительное количество верно предсказанных классов (доля верных ответов):

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}[y_{т_i} == y_i]$$

Пример.

Предсказанные	Реальные	$y_{т_i} == y_i$
1	0	0
0	0	1
1	1	1
0	0	1
1	0	0
0	0	1
0	0	1
1	1	1
0	0	1
0	0	1

$\sum_{i=1}^{10}[y_{т_i} == y_i] = 8$

1 - Редкий класс

0 - Частый класс

$Accuracy = 80\%$

Процедура обучения моделей на основе машинного обучения

Метрики качества модели классификации. Перекошенные классы

Accuracy – относительное количество верно предсказанных классов (доля верных ответов):

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}[y_{\text{т}i} == y_i]$$

Пример плохой модели с высокой долей верных ответов: $Y_{\text{т}}(X) = 0$

Предсказанные	Реальные
0	0
0	0
0	1
0	0
0	0
0	0
0	0
0	1
0	0
0	0

$Accuracy = 80\%$

	Реальный класс		
Предсказанный класс		«1» (*редкий класс)	«0»
	«1» (*редкий класс)	True positive (TP)	False positive (FP) Ошибка 1го рода
	«0»	False negative (FN)	True negative (TN)

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

Процедура обучения моделей на основе машинного обучения

Метрики качества модели классификации. Перекошенные классы

Матрица ошибок (confusion matrix). Точность и полнота модели классификации

	Реальный класс		
Предсказанный класс		«1» (*редкий класс)	«0»
	«1» (*редкий класс)	True positive (TP)	False positive (FP) Ошибка 1го рода
	«0»	False negative (FN) Ошибка 2го рода	True negative (TN)

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

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

$$F1-score = \frac{2PrecisionRecall}{Precision + Recall}$$

https://en.wikipedia.org/wiki/Confusion_matrix

Процедура обучения моделей на основе машинного обучения

Метрики качества модели классификации. Перекошенные классы

Матрица ошибок (confusion matrix). Чувствительность и специфичность модели

	Реальный класс		
Предсказанный класс		«1» (*редкий класс)	«0»
	«1» (*редкий класс)	True positive (TP)	False positive (FP) Ошибка 1го рода
	«0»	False negative (FN) Ошибка 2го рода	True negative (TN)

$$Sensitivity(Recall) = \frac{TP}{TP+FN}$$

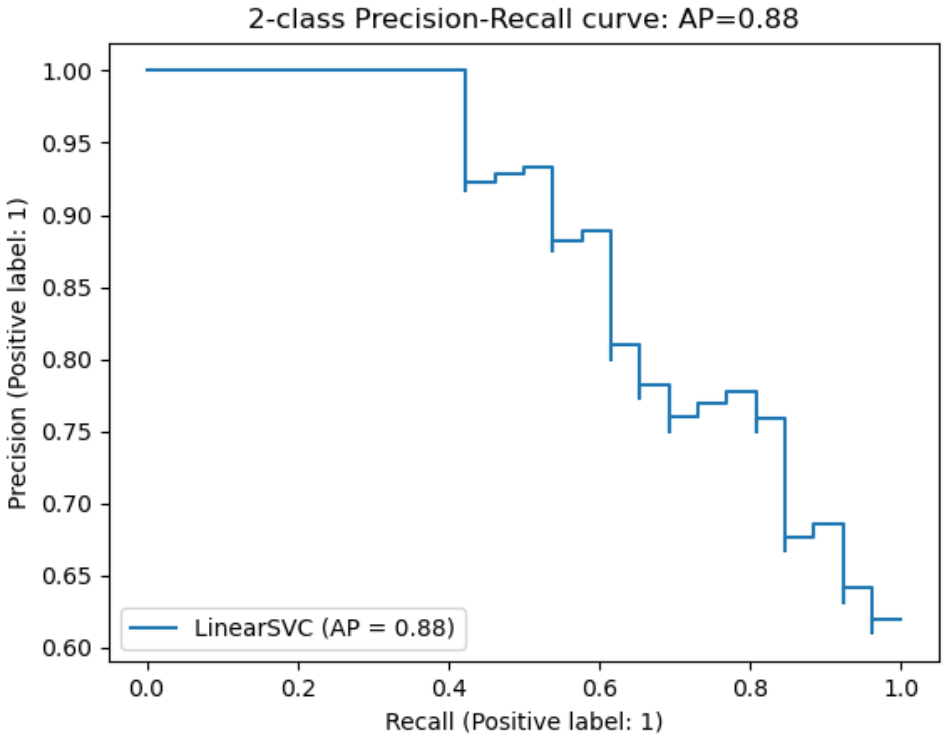
$$Specificity = \frac{TN}{TN+FP}$$

https://en.wikipedia.org/wiki/Confusion_matrix

Процедура обучения моделей на основе машинного обучения

Метрики качества модели классификации. Перекошенные классы

Precision/recall кривая



Классификатор: $y_{\text{т}i} = [h(X_i) > p]$

p	0	0,1	0,2	…	1
Precision					
Recall					

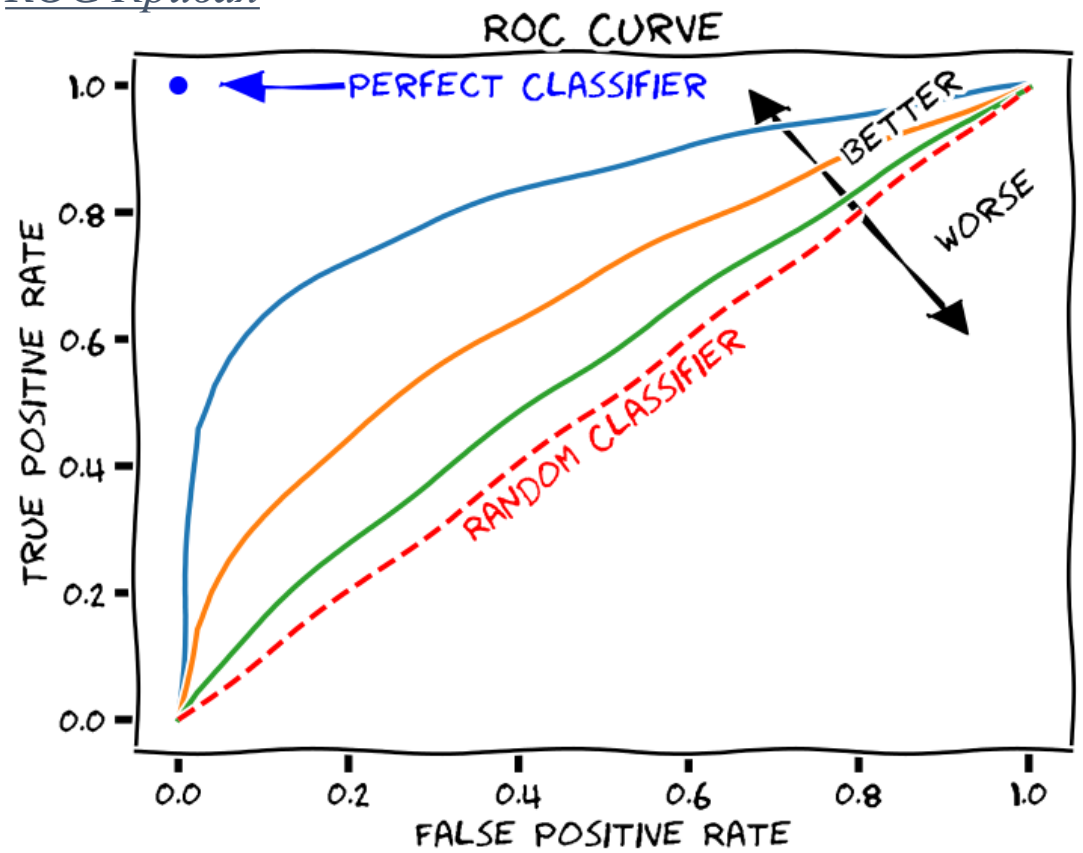
$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \ln(h_\theta(x_{ij})) + (1-y_i)(\ln(1-h_\theta(x_{ij}))\right) + \frac{\lambda}{2n}\sum_{j=1}^{k}\theta_j^2 \to \min$$

λ	0	0,01	0,1	1	10
Precision					
Recall					

Процедура обучения моделей на основе машинного обучения

Метрики качества модели классификации. Перекошенные классы

ROC Кривая



https://en.wikipedia.org/wiki/Receiver_operating_characteristic

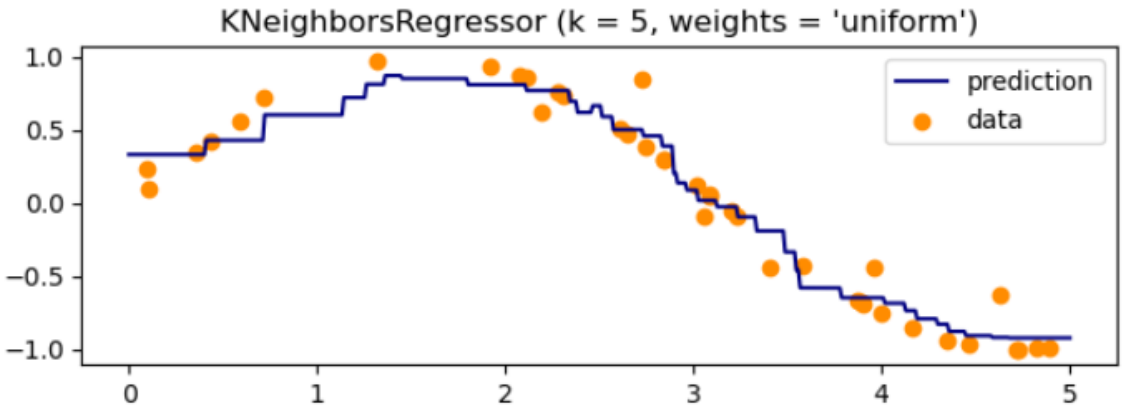
		Предсказанный класс	
		«1» (*редкий класс)	«0»
Предсказанный класс	«1» (*редкий класс)	True positive (TP)	False positive (FP) Ошибка 1го рода
	«0»	False negative (FN) Ошибка 2го рода	True negative (TN)

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{TP + FN}$$

Процедура обучения моделей на основе машинного обучения

Метод k-ближайших соседей для задачи регрессии

✓ Гипотеза компактности: предположение о том, что схожие объекты гораздо чаще лежат в одном классе, чем в разных



Обучение:

- Сохраняется обучающая выборка $\{X_i, Y_i\}$;

Предсказание отклика для новой точки:

- Измерить расстояние от всех объектов до нового объекта X_q;
- Упорядочить объекты в порядке возрастания дальности до нового объекта:

$$\rho(X_l, X_q) \leq \cdots \leq \rho(X_i, X_q) \leq \ldots \leq \rho(X_r, X_q)$$

- Выбрать первые k объектов (k ближайших соседей):

$$\{X_1, \ldots, X_k\}$$

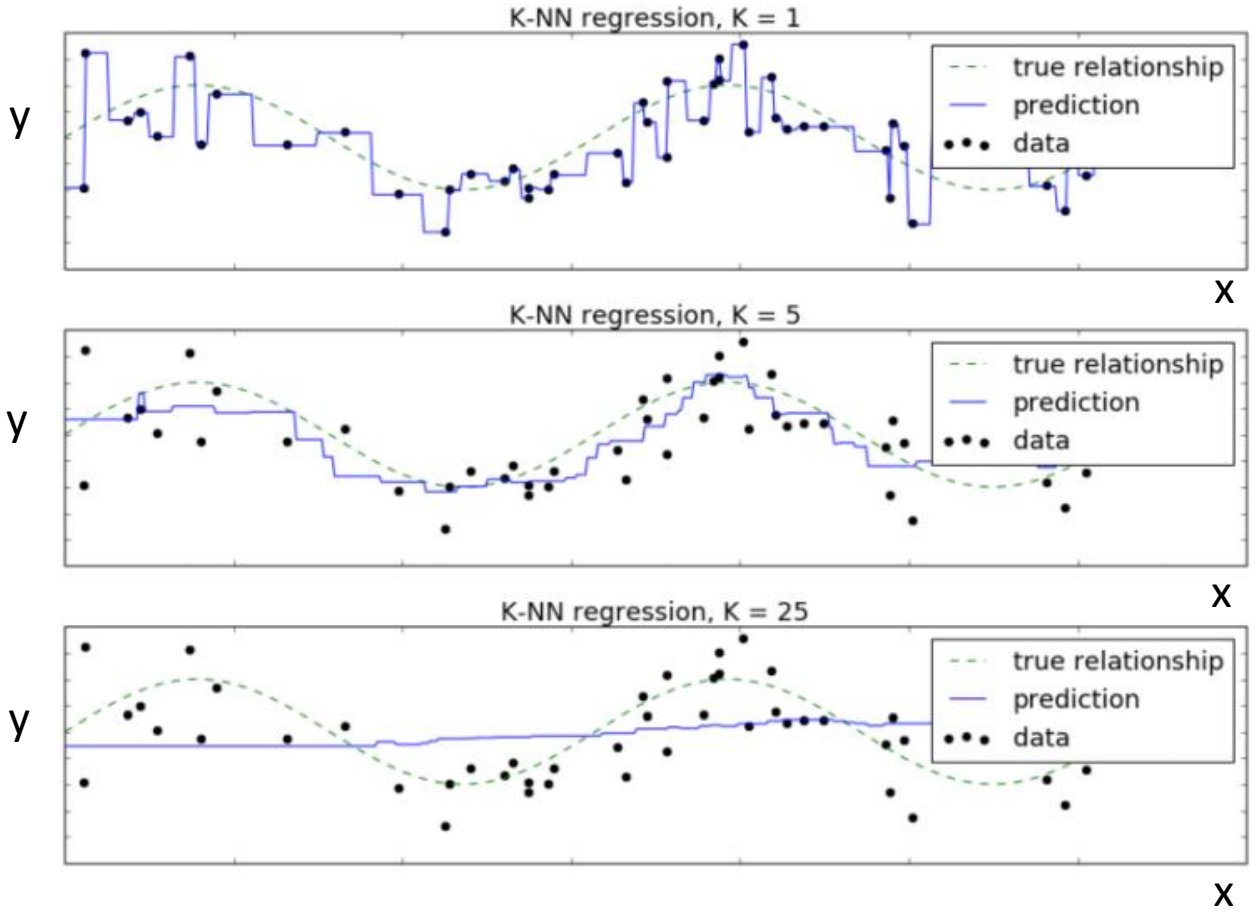
- Рассчитать среднее значение отклика среди k ближайших соседей:

$$Y(X_q) = \frac{1}{k}\sum_{i=1}^{k} y_i$$

Процедура обучения моделей на основе машинного обучения

Метод k-ближайших соседей для задачи регрессии

- Определение гиперпараметров на примере k-ближайших соседей



Валидация/кросс-валидация аналогично такой же процедуре для классификации, только рассматриваются метрики качества для регрессии.

Метрики качества для задач регрессии:

абсолютные:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y(x_i) - y_i)^2$$

$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y(x_i) - y_i|$$

относительные:

$$RSD = \frac{\sqrt{MSE}}{\bar{y}}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y(\vec{x}_i) - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$