Введение в АД

Лекция 2

Простейшие модели

ФЭФМ МФТИ

Весенний семестр 2023

Quick recap

- Features x, target y (features, target)
- model f (model)
- prediction $\hat{y} = f(x)$ (prediction)
- Loss function $L(y, \hat{y})$ (Loss function)
- sample $S = (x1, x2, \dots, xn)$
- hyperparameters

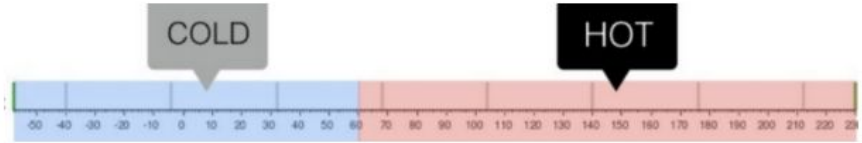
Regression

What will be the temperature tomorrow?

Fahrenheit

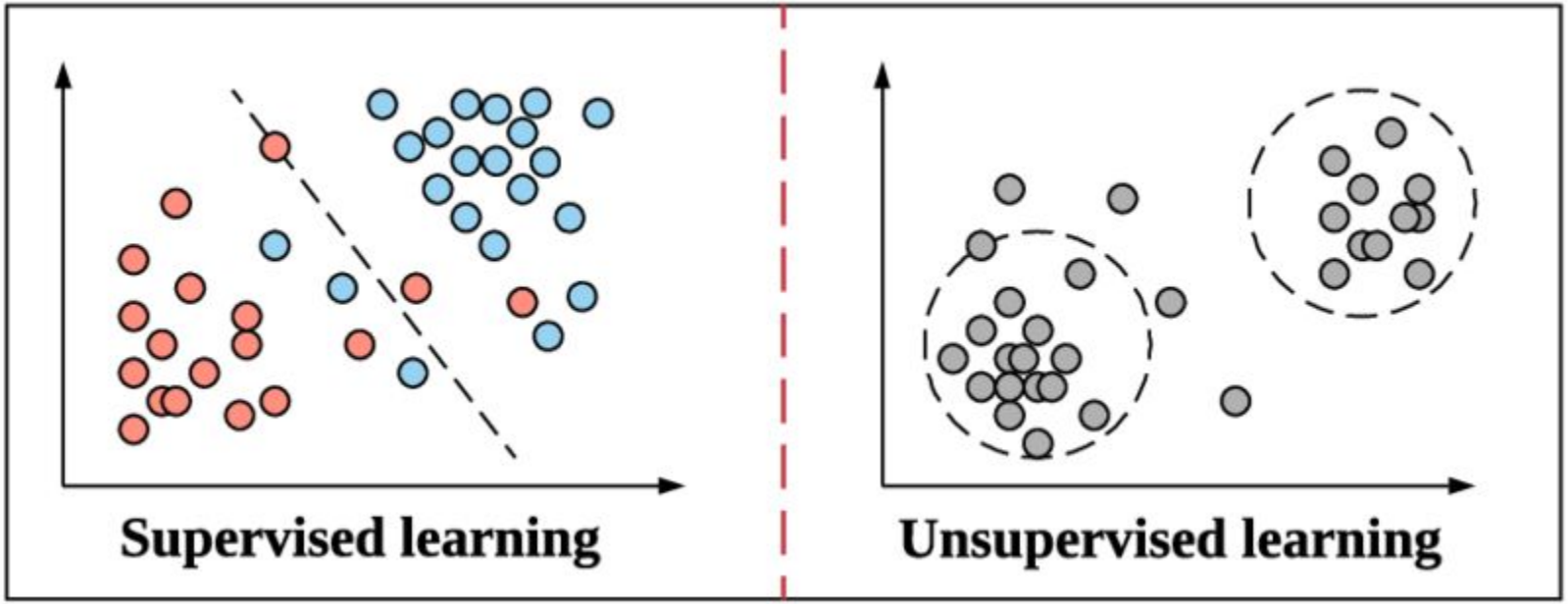
Classification

Will it be hot or cold tomorrow?

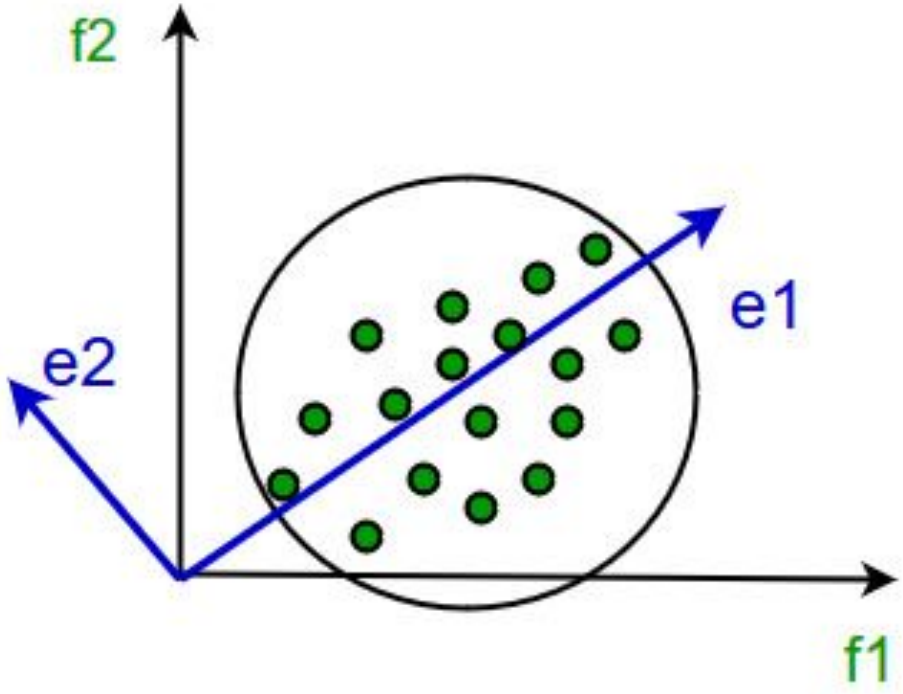


Fahrenheit

Не имеем изначальных классов, группируем точки по “похожести”



Имеем данные по принадлежности к красному или синему классу – предсказываем, что если новая точка слева от прямой, то она красная, если слева, то синяя



В этом случае, мы можем заменить представление точек в виде пары координат положением их проекции на e1, таким образом мы сократим размерность и сохраним наибольшее количество информации

Сегодня

- Naive Bayes classifier
- kNN algorithm
- Feature engineering
- Model life cycle
- Linear regression
- Regularization

Naive Bayes

Training set $\mathcal{L} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$, where

- $\mathbf{x}_i \in \mathbb{R}^p$, $y_i \in \{C_1, \ldots, C_k\}$ for k-class

Bayes theorem in this case:

$$\underbrace{P(y_i = C_k|\mathbf{x}_i)}_{\textit{posterior}} = \frac{\overbrace{P(\mathbf{x}_i|y_i = C_k)}^{\textit{likelihood}}\overbrace{P(y_i = C_k)}^{\textit{prior}}}{P(\mathbf{x}_i)}$$

The final class prediction:

$$C^* = \arg\max_k P(y_i = C_k|\mathbf{x_i})$$

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naive Bayes

Naive assumption – suggesting features are **mutually independent:**

$$P(\mathbf{x}_i|y_i = C_k) = \prod_{l=1}^{p} P(x_i^l|y_i = C_k)$$

How to get $P(x_i^l|y_i = C_k)$ in discrete and continuous cases?

$$\text{prior for a given class} = \frac{\text{no. of samples in that class}}{\text{total no. of samples}}$$

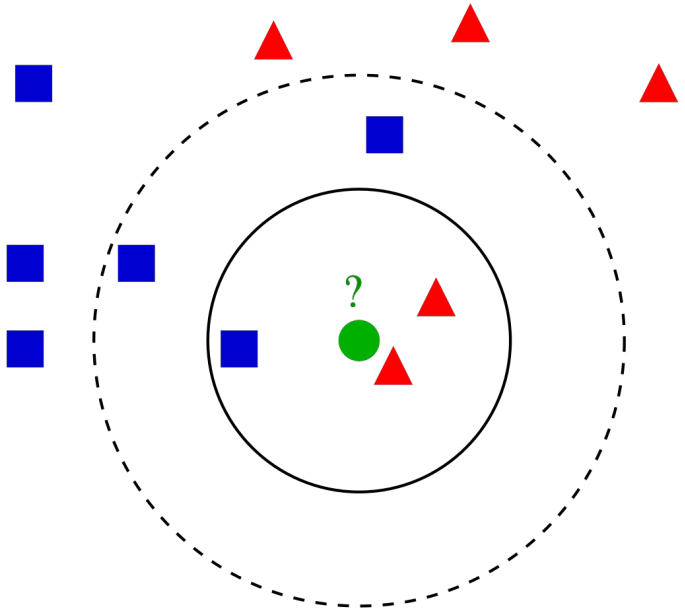
$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

kNN – k nearest neighbours

Non-parametric method

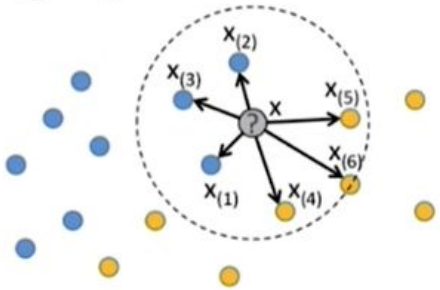
“Show me your friends, and I’ll tell you who you are”

Hyperparameter – number of neighbours to compare



Weighted kNN

k = 6



Weights can be adjusted according to the distance

$$w(\mathbf{x}_{(i)}) = w(d(\mathbf{x}, \mathbf{x}_{(i)}))$$

$$z_{\bullet} = \frac{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)})}{w(x_{(1)}) + w(x_{(2)}) + w(x_{(3)}) + w(x_{(4)}) + w(x_{(5)}) + w(x_{(6)})}$$

Can kNN be generalized to regression?

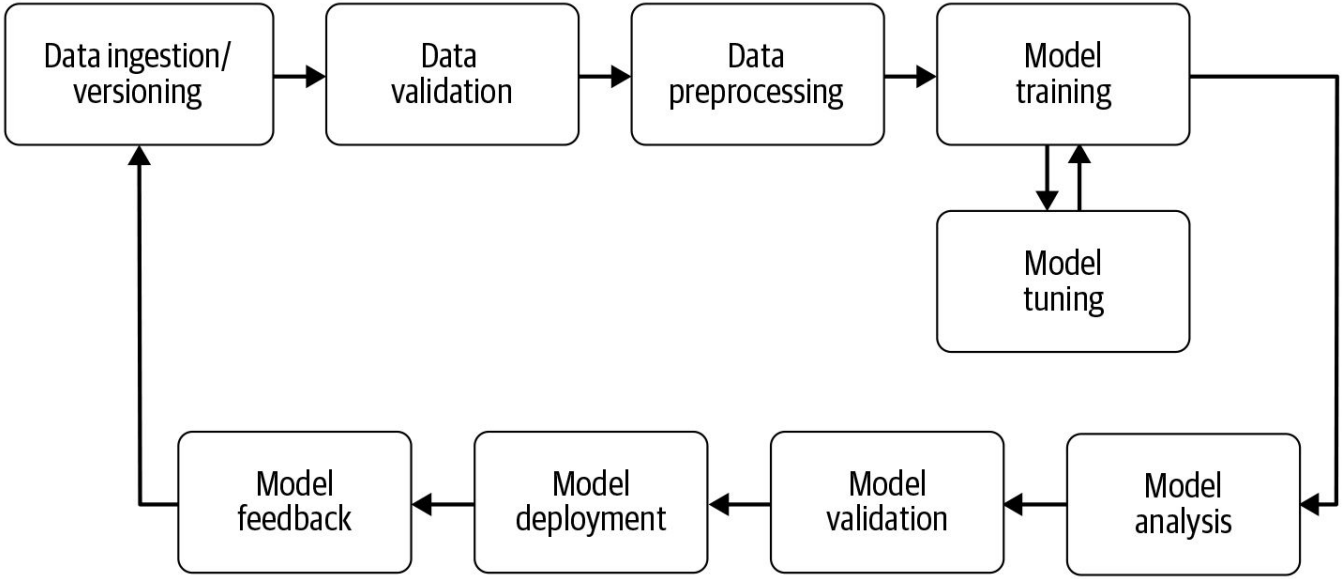
Feature engineering

	pet_type	color	weight
0	carrot	red	0.600000
1	cat	white	3.000000
2	hamster	brown	0.800000
3	cat	gray	5.000000
4	dog	black	7.000000

ДОРОГИЕ ШКОЛЬНИКИ! ПРИГЛАШАЕМ ВАС ПРИНЯТЬ УЧАСТИЕ В КОНКУРСЕ "БОЛЬШИЕ ВЫЗОВЫ" ! В ЭТОМ ГОДУ ФЭФМ МФТИ ПРОВОДИТ ТРЕК "ПЕРЕДОВЫЕ ПРОИЗВОДСТВЕННЫЕ ТЕХНОЛОГИИ" РАБОТЫ НА КОНКУРС ПРИНИМАЮТСЯ до 15 ФЕВРАЛЯ!

$$\mathbf{x}_i \in \mathbb{R}^p$$

Model life cycle



Linear regression

$$f_w(x_i) = \langle w, x_i \rangle + w_0$$

to exclude w0

$$(x_{i1} \quad \dots \quad x_{iD}) \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_D \end{pmatrix} + w_0 = (1 \quad x_{i1} \quad \dots \quad x_{iD}) \cdot \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$$

Let's suppose we have Mean-squared error (MSE) as a Loss function

$$\text{MSE}(f, X, y) = \frac{1}{N}|y - Xw|_2^2$$

how to differentiate matrix functions?

$$||Ax - b||^2 = \langle Ax - b, Ax - b \rangle$$

$$[D_{x_0}\langle Ax - b, Ax - b\rangle](h) =$$

$$\langle [D_{x_0}(Ax - b)](h), Ax_0 - b\rangle + \langle Ax_0 - b, [D_{x_0}(Ax - b)](h)\rangle$$

$$= 2\langle Ax_0 - b, [D_{x_0}(Ax - b)](h)\rangle =$$

$$= 2\langle Ax_0 - b, Ah\rangle = \langle 2A^T(Ax_0 - b), h\rangle$$

$$X^T(y - Xw) = 0$$

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

Gauss-Markov theorem

$$Y = f(X) + \varepsilon$$

$$\mathbb{E}(\varepsilon_i) = 0 \quad \forall i$$
$$\mathrm{Var}(\varepsilon_i) = \sigma_i^2 < \inf \quad \forall i$$
$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

Minimizing MSE loss gives

Best Linear Unbiased Estimation (BLUE)

(Estimator with minimal Variance from all unbiased estimators)

$$w^* = (X^T X)^{-1} X^T Y$$
$$\mathbb{E}(w^*) = w_{\text{true}}$$
$$Var(w^*) = min$$

NB: Matrix norm can be different

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2}.$$

Frobenius (Euclidean) norm

$$\|A\|_{\max} = \max\{|a_{ij}|\}$$

Max norm

Regularization

There are two problems regarding linear regression:

- Multicollinearity problem (features are nearly linear dependent, X^TX nearly singular)
- Low condition number (difference between eigenvalues of X^TX)

This result in unlikely big weights numbers in numerical solution

Thus we have regularization (L1, L2, or even both [ElasticNet])

$$L_2 = ||Y - Xw||_2^2 + \lambda^2||w||_2^2 \longrightarrow w = (X^TX + \lambda^2 I)^{-1}X^TY$$

$$L_1 = ||Y - Xw||_2^2 + \lambda^2||w||_1$$

$$MAE(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$MAPE(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

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

$$\text{SMAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \cdot |y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|}$$