Exercise 1.1

October 28, 2024

1 What are the stages of the ML cycle? Which ones are iterative stages?

The machine learning design cycle has the following phases:

- preprocessing
- feature extraction / encoding
- feature selection
- machine learning
- evaluation and model selection
- post-processing

Each step of this cycle is iterative, and revisited multiple times!

2 What are different types of learning?

There are multiple different types for ML. We distinguish between *supervised*, *semi-superivised*, *unsupervised* and *reinforcement*.

3 How would you describe the overfitting and underfitting phenomenon?

Underfitting means that our model, cannot represent the data, as the number of parameters is simply not enough for capturing the correct trend, or possibly cannot process irregularities in the data.

Overfitting means, that there are too many parameters in our model, that can be adjusted, and the model will simply learn the data, but cannot be used to make further predictions, with data points, that were not learned before.

Exercise 1.2

October 30, 2024

1 Estimate the probabilities

$$p(\text{yes}) = \frac{6}{10} \tag{1}$$

$$p(\text{red} \mid \text{yes}) = \frac{3}{6} = \frac{1}{2} \tag{2}$$

$$p(\text{grand tourer} \mid \text{yes}) = \frac{2}{6} \tag{3}$$

$$p(\text{domestic} \mid \text{yes}) = \frac{2}{6} \tag{4}$$

$$p(\text{no}) = \frac{4}{10} \tag{5}$$

$$p(\text{red} \mid \text{no}) = \frac{1}{4} \tag{6}$$

$$p(\text{grand tourer} \mid \text{no}) = \frac{2}{4} \tag{7}$$

$$p(\text{domestic} \mid \text{no}) = \frac{3}{4} \tag{8}$$

2 Predict the probability that a car with properties x_1 = red, x_2 = grand tourer, x_3 = domestic will be stolen.

$$p(\text{yes}) \cdot p(\text{red} \mid \text{yes}) \cdot p(\text{grand tourer} \mid \text{yes}) \cdot p(\text{domestic} \mid \text{yes})$$
$$= \tfrac{6}{10} \cdot \tfrac{1}{2} \cdot \tfrac{2}{6} \cdot \tfrac{2}{6}$$
$$= \tfrac{3}{40}$$

3 What are the benefits, what are the downsides of using Naive Bayes?

The benefits are that it does not require a lot of data and is easy to implement, being ideal for smaller datasets. It is also quite fast.

The downside is that it is too dependent on data being of high quality/ Noisy data can result in incorrect results. It also assumes that all features are independent, which can limit its effectiveness where this is not the case.

4 The extra mile: Derive Equation 1 using Bayes' theorem, the chain rule of probabilities and the conditional independence assumption stated above.

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Chain Rule of Probability:

$$P(A \wedge B) = P(A) \cdot P(B)$$

Conditional Independence Assumption:

$$P(A \wedge B|C) = P(A|C) \cdot P(B|C) \quad \text{if A and B are conditionally independent given C}$$

Goal: Estimating conditional probability of y

$$P(y = k|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|y = k) \cdot P(y = k)}{P(x_1, x_2, \ldots, x_n)}$$

Expanding the numerator using conditional independence:

$$= \frac{P(x_1|y = k) \cdot P(x_n|y = k)}{P(x_1, x_2, \ldots, x_n)} \cdot P(y = k)$$

Using the product rule for conditional probabilities:

$$= \frac{1}{Z} P(y = k) \prod_{i=1}^{n} P(x_i|y = k)$$

with $Z = P(x) = P(x_1, x_2, \ldots, x_n)$.

Exercise 1.3

October 30, 2024

1 Formulate mathematically a loss function

$$\sum_{i<j} 1(y_i > y_j)$$

Boolean statements will evaluate to either 1 (True) or 0 (False). This loss function checks that each input is higher ranked than the following. Any correct rankings do not contribute to the loss of value, whereas incorrect ones do.

2 Which model is better at ranking? Why is the squared error problematic in this case?

The tuples that meet the condition $i < j$ are $(1, 2)$, $(1, 3)$ and $(2, 3)$, so these will be used in the summation.

For the first model $\hat{y}_1$, the loss function evaluates to the following.

$$\begin{aligned} & 1(1 > 3) + 1(1 > 2) + 1(3 > 2) \\ & = 0 + 0 + 1 \\ & = 0 \end{aligned} \tag{1}$$

For the second model $\hat{y}_2$, the loss function evaluates to the following.

$$\begin{aligned} & 1(2 > 3) + 1(2 > 7) + 1(3 > 7) \\ & = 0 + 0 + 0 \\ & = 0 \end{aligned} \tag{2}$$

So the second model $\hat{y}_2$ is better. Squared error is problematic because it would penalise $\hat{y}_2$ more for having a larger range despite it having a more correct ranking.