ID2223 - Scalable Machine Learning and Deep Learning

- Review Questions 1 -

PAOLO TETA RALFS ZANGIS

teta | zangis @kth.se

November 12, 2021

1 Question 1

(a) T

(b) T

(c) F

2 Question 2

Squared error: $(-0.2)^2 + (0.4)^2 + (-0.8)^2 + (1.3)^2 + (-0.7)^2 = 0.04 + 0.16 + 0.64 + 1.69 + 0.49 = 3.02$
Mean squared error (MSE): $3.02/5 = 0.604$

3 Question 3

Correct answers: A, D.

4 Question 4

A simple linear regression model needs only 2 coefficients, a and b, when it has one independent variable. Example linear regression function: $y = ax + b$. In this equation, a is the coefficient, x is the predictor (input), and b is the constant parameter (bias).

5 Question 5

Cross-validation is a technique to asses the performance of models while avoiding using excessive amounts of training data in validation sets. In cross-validation, the solution is tested by creating a certain number of folds/partitions (k-fold cross-validation) from the available dataset. Then, the model is trained and tested using an ever-changing selection of partitions result, providing in the end the overall error. In detail, firstly we train the model against a batch of folds and then the model is validated against the remaining ones.

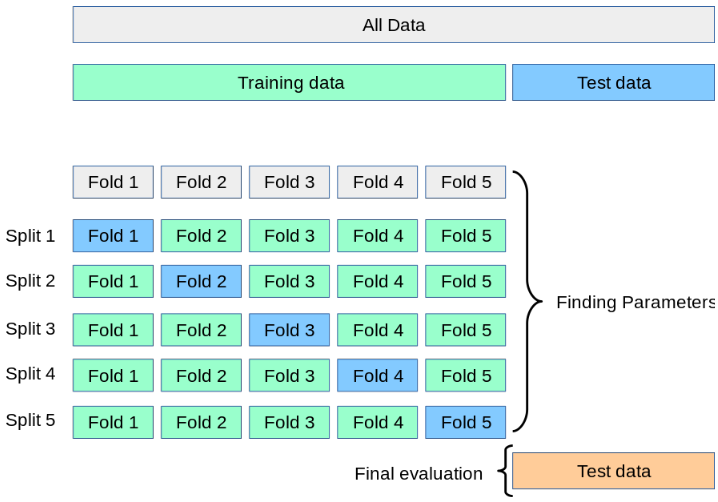

Figure 1: Cross-validation technique example (k = 5)

6 Question 6

Given the following functions:

- sigmoid function $\longrightarrow \sigma(w^\intercal x) = \dfrac{1}{1+e^{-w^\intercal x}}$
- softmax function $\longrightarrow \sigma(w_j^\intercal x) = \dfrac{e^{w_j^\intercal x}}{\sum_{i=1}^{k} e^{w_i^\intercal x}}$

The sigmoid function is used for the two-class logistic regression, whereas the softmax function is used for the multiclass logistic regression. The softmax function is an extension of the sigmoid function to the multiclass case using $K = 2$ as the total number of classes (example below).

$$p(y=0|x;w) = \sigma(w^\intercal x) = \frac{e^{-w^\intercal x}}{1+e^{-w^\intercal x}} \qquad \textit{and} \qquad p(y=1|x;w) = \sigma(w^\intercal x) = \frac{1}{1+e^{-w^\intercal x}}$$

$$p(y=0|x;w_0) = \sigma(w_j^\intercal x) = \frac{e^{w_j^\intercal x}}{\sum_{i=0}^{k} e^{w_i^\intercal x}} = \frac{e^{w_0^\intercal x}}{e^{w_0^\intercal x}+e^{w_1^\intercal x}} = \frac{e^{(w_0^\intercal - w_1^\intercal)x}}{e^{(w_0^\intercal - w_1^\intercal)x}+1} = \frac{e^{-w^\intercal x}}{1+e^{-w^\intercal x}}$$
$$p(y=1|x;w_1) = \sigma(w_j^\intercal x) = \frac{e^{w_j^\intercal x}}{\sum_{i=0}^{k} e^{w_i^\intercal x}} = \frac{e^{w_1^\intercal x}}{e^{w_0^\intercal x}+e^{w_1^\intercal x}} = \frac{1}{e^{(w_0^\intercal - w_1^\intercal)x}+1} = \frac{1}{1+e^{-w^\intercal x}}$$

with $w^\intercal = -(w_0^\intercal - w_1^\intercal)$.
So, as we can see the softmax function with two classes (K = 2) is equivalent to the sigmoid function.

7 Question 7

In logistic regression, $-log(x)$ function is a proper function to compute the cost because it produces a huge number when x is getting close to 0, while being 0 when x is close to 1. So, in the implementation the cost would be high when a wrong probability is estimated.

8 Question 8

The negative log-likelihood is equal to the cross-entropy. The latter quantifies the difference between two probability distributions. The cross-entropy of true and predicted probabilities is equal to the logistic regression cost. So, they are all related via the maximum likelihood estimation principle.

9 Question 9

The ROC curve illustrates the predictive performance of a model. This is done by plotting the true positive rate (TPR), known as sensitivity, against false positive rate (FPR). The better the model performs, the closer the curve should be to the left upper corner of the plot.

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