KTH Royal Institute of Technology
Stockholm

School of Electrical Engineering and
Computer Science

Scalable Machine Learning and Deep Learning - ID2223

Review Questions 1

Author
Emil Ståhl

Author
Selemawit Fsha

Author
Erik Kongpachith

November 12, 2021

Review Questions 1 - ID2221

Emil Ståhl, Erik Kongpachith, and Selemawit Fsha Nguse

November 12, 2021

1 Which of the following is/are true about Normal Equation?

(a) We don't have to choose the learning rate.

True, no need to choose learning rate.

(b) It becomes slow when number of features is very large.

True, generally if the number of features is less than 10000, one can use normal equation to get the solution beyond which the order of growth of the algorithm will make the computation very slow. Normal equation works well with small number of features.

(c) No need to iterate.

True, with Normal Equation we do not need to iterate, we get a direct solution. It is an analytical approach.

2 The following graph represents a regression line predicting y from x. The values on the graph shows the residuals for each predictions value, i.e., $\hat{y}$ - y. Calculate the squared error of the prediction.

Squared error =

$$-0,2^2 + 0,4^2 + -0,8^2 + 1.3^2 + -0,7^2 = 3,02$$

3 How does number of observations influence overfitting? Choose the correct answer(s).

(a) In case of fewer observations, it is easy to overfit the data.

Answer: **True**

(b) In case of fewer observations, it is hard to overfit the data.

(c) In case of more observations, it is easy to overfit the data.

(d) In case of more observations, it is hard to overfit the data.

Answer: **True**

(a) and (d) are true

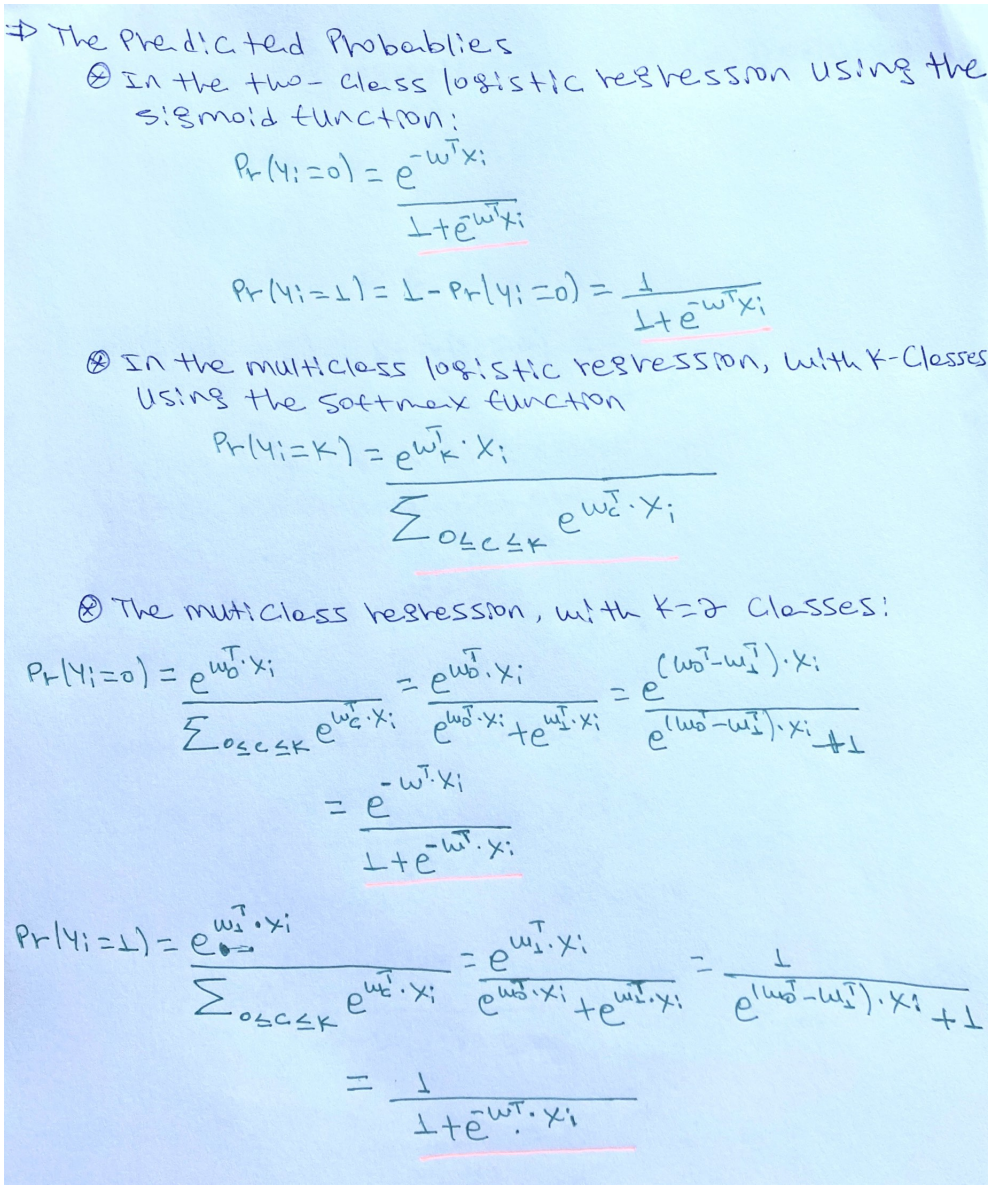
4 How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

In simple linear regression, there is one independent variable so **2** coefficients (Y=a+bx) are needed. Number of independent variables + 1.

5 What is cross validation and how does it work?

Cross-validation is a technique used to protect against overfitting in a predictive model, particularly in a case where the amount of data may be limited. In cross-validation, you make a fixed number of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate.

6 Mathematically show that the softmax function with two classes (k = 2) is equivalent to the sigmoid function?



7 As you know, in binomial logistic regression the cost between the true value y and the predicted value ŷ is measured as below: Explain why -log is a proper function to compute the cost in logistic regression?

Using the same definition for the cost function as in linear regression would result in a non-convex cost function, which means that a local minimum can be found before reaching the global minimum. To ensure convergence to the global minimum, the -log function is used instead to also ensure that the two following conditions are met:

cost($\hat{y}$, y) is

- Close to 0, if the predicted value $\hat{y}$ will be close to true value y.
- Large, if the predicted value $\hat{y}$ will be far from the true value y

8 How are logistic regression cost, cross-entropy, and negative log-likelihood related?

The logistic regression cost, cross-entropy and negative log-likelihood can be derived from the same equation that is:

$$-\frac{1}{m}\sum_{i}^{m}(y^{(i)}\log(\hat{y}^{(i)}) + (1-y^{(i)})\log(1-\hat{y}^{(i)}))$$

9 Explain how a ROC curve works?

A ROC curve displays the trade off between the true positive rate(TPR) and false positive rate(FPR) for a modeled classifier. Classifiers which have a ROC curve closer to the top-left corner of a ROC curve plot indicate a better performance. One can measure the performance between different classifiers with a single measure using the area under the curve of the ROC curve. Lowering the classification threshold would classify more inputs to True Positives, thus increasing both TPR and FPR. This could be evaluated as a performance measure of three different classifiers at various classification thresholds. It would be really inefficient calculate the ROC values when we deal with millions of data points. One solution for this is called Area Under the Curve or AUC Curve. ROC is a probability curve and AUC represents the degree or measure of separability between classes. We conclude, higher the AUC, better the model.