Scalable Machine Learning and Deep Learning
Research Questions 1

Boffo Marco
Dei Rossi Marco

November 6, 2020

Abstract

In the following document, you can find the answers to the Research Questions 1 Assignment.

QUESTION 1 (0.5 point)

Which of the following are true about *Normal Equation*?

All of them are true:
a) We don't have to choose the learning rate.
b) It becomes slow when the number of features is very large.
c) No need to iterate

QUESTION 2 (0.5 point)

Calculate the squared error of the prediction.

Using the formula for the squared error: $SE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$,
we obtain as result $SE = 3.02$.

QUESTION 3 (0.5 point)

How does number of observations influence overfitting?

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

QUESTION 4 (0.5 point)

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

The simple linear regression is defined from a simple equation of a line in a plane with one dependent and one independent coordinate, as such we need 2 coefficient to obtain: $\hat{y} = b + wx$, where b is called *intercept/bias* and w is called *slope*.

QUESTION 5 (0.5 point)

What is cross validation and how does it work?

Cross-validation is a technique to avoid wasting too much training data in the validation sets, by using a re-sampling procedure to evaluate training models on a limited amount of data sample.

It is used to estimate how the model is expected to perform in general when used to make predictions on data not used in the training model.

If we have a small data-set, we cannot cut out data from the training process because we could have a problem of underfitting, so the initial training set will be split into actual training and validation subsets.

Each model is trained against a different combination of these subsets and validated against the remaining parts. The final training error is averaged over all the trials.

Once the model type and hyperparameters have been selected, the final model is trained using these hyperparameters on the full training set, and the test error is measured on the test set.

QUESTION 6 (0.5 point)

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

From the definition of:

- Sigmoid function: $\hat{y} = p(y = 1|x; w) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$;
- Softmax function: $\hat{y_j} = p(y = j|x; w_j) = \sigma(w_j^T x) = \frac{e^{w_j^T w}}{\sum_i^k e^{w_i^T w}}$;

we can say that for binary class ($y = 0, 1$) / two classes ($k = 2$) the functions are equivalent.

The Sigmoid present the probability that y have value equal to 1:

$$\hat{y} = p(y = 1|x; w) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}} = \frac{e^{w^T x}}{1 + e^{w^T x}}$$

The probability of having value $y = 0$, in this binary setting, is therefore $1 - \hat{y}$:

$$!\hat{y} = 1 - \hat{y} = 1 - p(y = 1|x; w) = 1 - \sigma(w^T x) = 1 - \frac{1}{1 + e^{-w^T x}} = \frac{1 + e^{-w^T x} - 1}{1 + e^{-w^T x}} = \frac{e^{-w^T x}}{1 + e^{-w^T x}}$$

The Softmax can be manipulated in this way:

$$\hat{y_1} = p(y = 1|x; w_1) = \sigma(w_1^T x) = \frac{e^{w_1^T w}}{\sum_i^k e^{w_i^T w}} = \frac{e^{w_1^T x}}{e^{w_1^T x} + e^{w_0^T x}}$$

$$\hat{y_0} = p(y = 0|x; w_0) = \sigma(w_0^T x) = \frac{e^{w_0^T w}}{\sum_i^k e^{w_i^T w}} = \frac{e^{w_0^T x}}{e^{w_1^T x} + e^{w_0^T x}}$$

Setting w_1 such that $w_1^T x = 0$ and $w0$ such that $w_0^T x = -w^T x$, we can conclude that the values obtained using the Sigmoid $\hat{y}$ and $!\hat{y}$ are the same as the values obtained using the Softmax function $\hat{y_1}$ and $\hat{y_0}$, respectively.

QUESTION 7 (0.5 point)

Explain why $-log$ is a proper function to compute the cost in logistic regression?

We want the obtain the values of the cost function $cost(\hat{y_i}, y_i)$:

- Close to 0, if the predicted value $\hat{y}$ is close to true value y;
- Exponentially large, if the predicted value $\hat{y}$ is far from the true value y;

So, we use a $-log$ function to ensure that the error grows exponentially with respect to the distance of the predicted value from the true one.

QUESTION 8 (0.5 point)

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

The **logistic regression cost function** is the following:

$$J(\mathbf{w}) = -\frac{1}{m}\sum_{i}^{m}(y_i * log(y_i) + (1 - y_i) * log(1 - y_i))$$

The **negative log-likelihood function** instead is the following:

$$-log(L(\mathbf{w})) = -\sum_{i}^{m} log(p(y_i|\mathbf{x}_i; \mathbf{w})) = -\sum_{i}^{m}(y_i * log(y_i) + (1 - y_i) * log(1 - y_i))$$

We can then say that these are directly linearly related as such: $J(\mathbf{w}) = \frac{1}{m} * (-log(L(\mathbf{w})))$.
Therefore, minimizing the negative log-likelihood will also minimize the cost function $J(\mathbf{w})$.

Sigmoid cross-entropy quantifies the difference (error) between two probability distributions:

$$H(p, q) = \sum_{j} p_j log(q_j)$$

Setting p as true distribution (so that: $p(y = 1) = y$ and $p(y = 0) = 1 - y$) and q as predicted distribution (so that: $q(y = 1) = \hat{y}$ and $q(y = 0) = 1 - \hat{y}$), we obtain the following:

$$H(p, q) = -\sum_{j} p_j log(q_j) = -(y * log(\hat{y}) + (1 - y) * log(1 - \hat{y}))$$

We can then conclude saying that the logistic regression cost function and the cross-entropy are directly related as: $J(\mathbf{w}) = \frac{1}{m}\sum_{j=1}^{m} H(p, q)$.
So, minimizing the cross-entropy will also minimize the cost function $J(\mathbf{w})$.

QUESTION 9 (0.5 point)

Explain how a ROC curve works?

The Receiver Operating Characteristic (ROC) curves is a curve that shows the performance of a model at a classification threshold. To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification/probability thresholds, summarizing the trade-off between the True Positive Rate ($p(\hat{y} = 1|y = 1)$) and False Positive Rate ($p(\hat{y} = 1|y = 0)$) for a model. Generally, the higher the TPR, the more FPR the classifier produces.

A good Classifier place the curve close to the top left corner, or rather with large True Positive Rate and small False Positive Rate.