08 November 2020

Scalable Machine Learning and Deep Learning - Review Questions 1

By

Akhil Yerrapragada (akhily@kth.se)

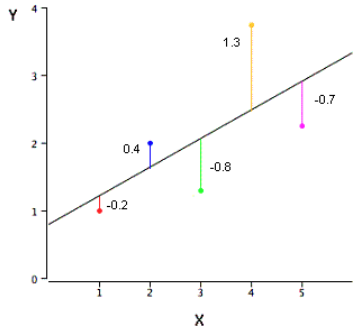
Gibson Chikafa (chikafa@kth.se)

1. **0.5 point.** Which of the following is/are true about *Normal Equation*?
   (a) We don't have to choose the learning rate.
   (b) It becomes slow when number of features is very large.
   (c) No need to iterate.

Ans: a, b, c

---

2. **0.5 point.** The following graph represents a regression line predicting yfrom **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.



Ans: Squared Error (SE) = (-0.2)^2 + (0.4)^2 + (-0.8)^2 + (1.3)^2 + (-0.7)^2 = 3.02

---

3. **0.5 point.** 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.
   (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.

Ans: a, d

---

4. **0.5 point.** How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

Ans: 2 (m (Slope) and c (Y - Intercept) in y = mx + c)

---

5. **0.5 point.** What is cross validation and how does it work?

Ans: It is an approach to choose between training and validation dataset in training dataset. Here, we split training data into n subsets. The model is trained with each n per iteration and is validated with other parts. The achieved average of iterations at the end will be the values we can use as hyperparameters.

---

6. **1 point.** Mathematically show that the softmax function with two classes (k = 2) is equivalent to the sigmoid function?

Proof:

We know that Sigmoid for $y \in \{0,1\}$ is

$$\hat{y} = P(y=1 \mid x; w) = \frac{1}{1+e^{-w_1^T x}}$$

$$\Rightarrow \frac{1}{1+\frac{1}{e^{w_1^T x}}}$$

$$P(y=1) \Rightarrow \frac{e^{w_1^T x}}{1+e^{w_1^T x}} \quad \text{— (1)}$$

Now,

let us Consider 2 class softmax function for $y \in \{4,5\}$

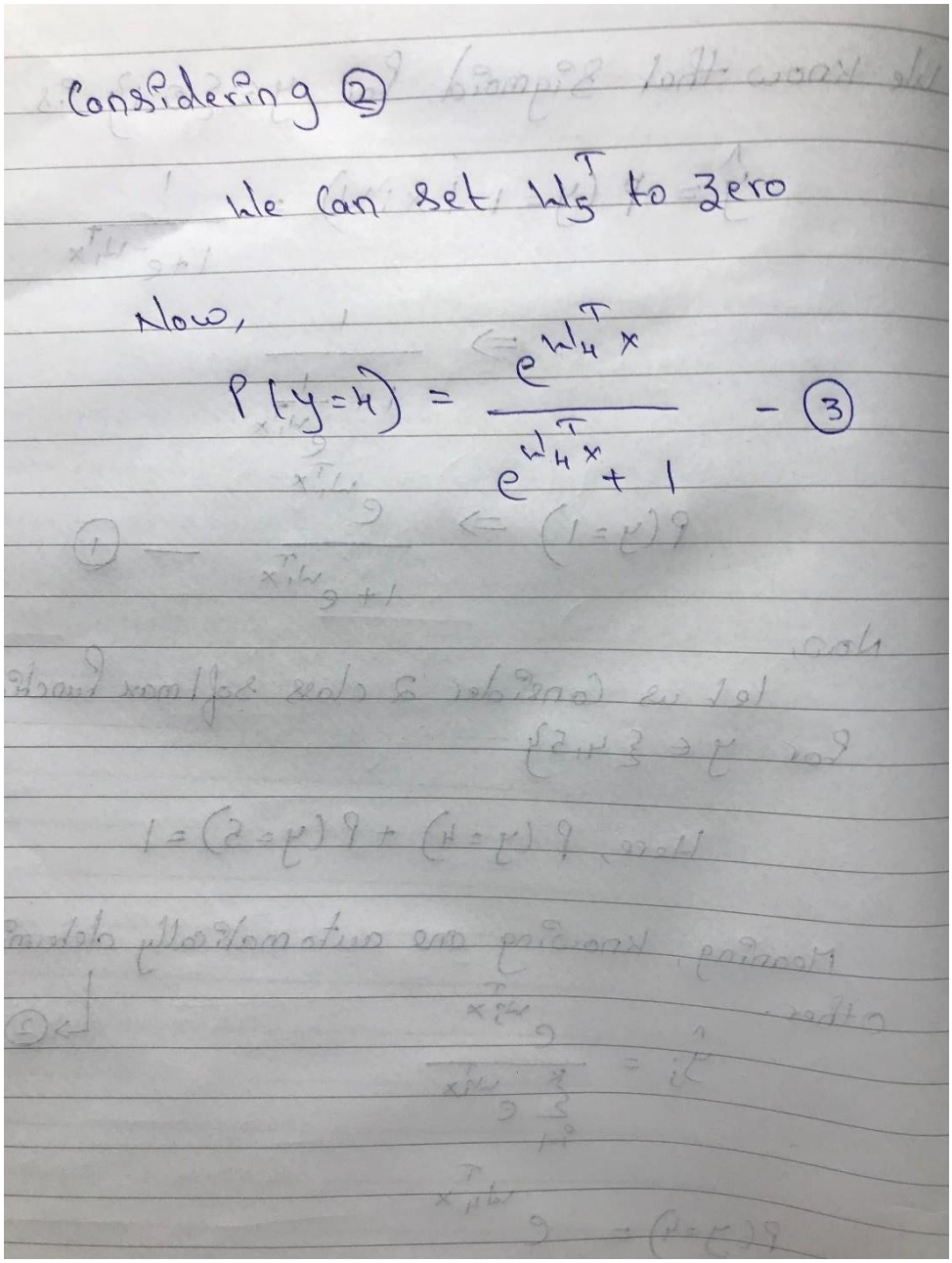
Here, $P(y=4) + P(y=5) = 1$

Meaning, knowing one automatically determines other. → (2)

$$\hat{y}_j = \frac{e^{w_j^T x}}{\sum_{i=1}^{k} e^{w_i^T x}}$$

$$P(y=4) = \frac{e^{w_4^T x}}{e^{w_4^T x} + e^{w_5^T x}}$$

08 November 2020



Considering 1 and 3 to be similar in the above images, we can say that 2 class Softmax is equivalent to Sigmoid function. We can also consider the Softmax for y = {0, 1} which would yield same results. Here, since we know that p(4) + p(5) = 1, knowing one would automatically yield the result for other. Therefore, we can substitute in the equation W5 to be zero.

7. **0.5 point.** As you know, in binomial logistic regression the cost between the true value y and the predicted value ŷ is measured as below:

$$\text{cost}(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & \text{if } y = 1 \\ -\log(1-\hat{y}) & \text{if } y = 0 \end{cases}$$

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

Ans: This is because -log gives high cost when true value is not similar to predicted value and zero cost when true value is similar to predicted value.

For example:

1) **If actual(y) = 1 and predicted value(y^) = 1 we want the cost to be zero.**

Applying -log to the predicted gives zero cost as shown below:

⇨ **-log(1) = 0**

2) **If actual(y) = 1 and predicted value(y^) = 0 we want the cost to be high.**

⇨ **-log(0) = infinity**

3) **If actual(y) = 0 and predicted value (y^) = 0 we want the cost to be zero.**

⇨ **-log(1-0) = 0**

4) **If actual(y) = 0 and predicted value (y^) = 1 we want the cost to high.**

⇨ **-log(1-1) = infinity**

The negative symbol is because log(0) gives -ve infinity and cost cannot be -ve. -log(0) gives infinity. Hence -log is a proper function.

---

8. **0.5 point.** How are logistic regression cost, cross-entropy, and negative log-likelihood related?

Ans: Negative log likelihood is used to quantify the difference between two probability distributions. It is also called as Cross-Entropy.

To minimize cost function J(w), minimize negative log likelihood. Also, to minimize cost function J(w), minimize cross-entropy. Below is the explanation:

The negative log likelihood is given by the following equation:

$$-\log(L(w)) = -\sum_{i=1}^{m} y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \quad \text{< Equation 1 >}$$

The cost function J(w), can be given by:

$$J(w) = -\frac{1}{m} \sum_{i} (y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})) \quad \text{< Equation 2 >}$$

The cross entropy of p and q is given by:

$$H(p, q) = -\sum_{j} p_j \log(q_j) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) = \text{cost}(y, \hat{y}) \quad \text{< Equation 3 >}$$

By comparison between Equation 2 and Equation 3 we can observe that they are similar:

J(w) = 1 m Xm i cost(y, ^y) = 1 m Xm i H(p, q) = − 1 m Xm i (y (i) log(^y (i)) + (1 − y (i))log(1 − ^y (i)))

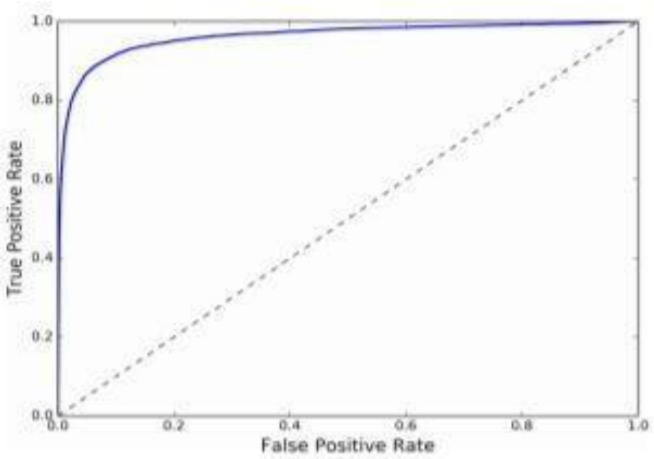
Therefore, to minimize cost function J(w), minimize cross-entropy.

Similarly, comparison between Equation 1 and Equation 2 states that, to minimize cost function J(w), minimize negative log likelihood since they are similar.

---

9. **0.5 point.** Explain how a ROC curve works?

Ans:



KTH – Course ID2223 – Machine Learning Classification - Lecture slide 67

ROC (Receiver operating characteristic) curve shows the relation between Precision and Recall taking into consideration of different thresholds. The X axis of ROC curve is False Positive Rate (FPR), meaning, it denotes Precision. Y axis is True Positive Rate (TPR), meaning, it denotes Recall. ROC portrays the performance of a classifier. Classifier to be considered good, it must move towards Top-Left. For it to be considered perfect classifier, value at Y must be 1 and value at X must be zero.

True Positive Rate is given by:

TPR = TP / (TP + FN)

False positive rate is given by:

FPR = FP / (FP + TN)