Quiz Submissions - Quiz 5

Mike Gao (username: zenghao.gao@mail.mcgill.ca)

Attempt 1

Written: Oct 19, 2020 12:19 AM - Oct 19, 2020 7:28 PM

Submission View

Released: Oct 18, 2020 11:26 AM

Logistic Regression

Question 1 — 1 / 1 point

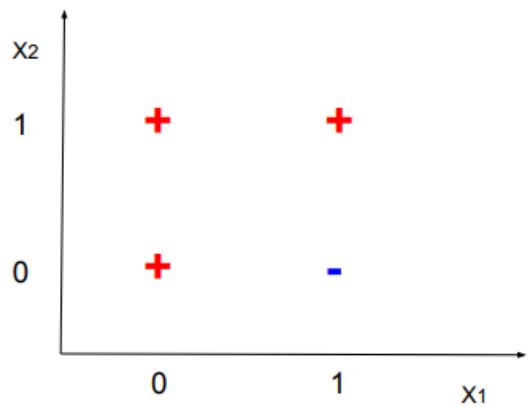
Suppose that you trained a logistic regression classifier h_w for a binary prediction task. When you apply it on a new sample x', you get a prediction h_w(x)=0.7. It means (check all that apply):

- [ ] **P(y=0|x;w)=0.7**
- [x] **P(y=1|x;w)=0.7**
- [x] **P(y=0|x;w)=0.3**
- [ ] **P(y=1|x;w)=0.3**

Question 2 — 1 / 1 point

Suppose we have two features with boolean variables X1, X2 in {0,1} and label Y in {0,1} where "+" sign below denotes Y=1 and "-" sign denotes Y=0.

In the image below, can logistic regression perfectly classify the examples?



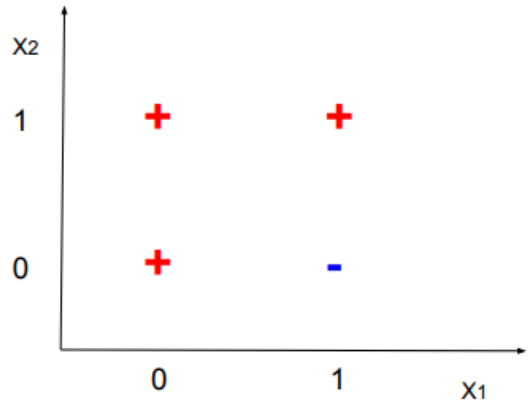
- (•) True
- ( ) False

Hide Feedback

Because the data set in the above image is linearly separable, Logistic Regression would be perfectly able to classify them.

Question 3 **1 / 1 point**

In the image, if we change the label of point (0,1) to "-" label. Would logistic regression be able to perfectly classify the examples?



- True
- ✓ False

Hide Feedback

Data points are not linearly separable, so LR would not be able to perfectly classify the above examples.

Question 4 **1 / 1 point**

The decision boundary of both logistic regression and k-NN is always linear.

- True
- ✓ False

Hide Feedback

Decision boundary of logistic regression is always linear, while for kNN it can be non-linear.

Linear Regression

Question 5 — 0 / 1 point

The maximum likelihood model with parameters (w) can be learned using linear regression for the model:

$$y_i = \log(x^{w_1} e^{w_2}) + \epsilon_i$$

where

$$\epsilon_i = N(0, \sigma^2)$$

iid noise and x>0 is a feature.

- True
- ✖ False

Hide Feedback

$$\log(x^{w_1} e^{w2}) + \epsilon_i = w_1 \log(x) + w_2 + \epsilon_i$$

which is still linear in parameter w_1 and w_2.

Question 6 — 1 / 1 point

The maximum likelihood model with parameters (w) can be learned using linear regression for the model:

$$y_i = w.\, x_1 x_2^3 + \epsilon_i$$

, where

$$\epsilon_i = N(0, \sigma^2)$$

iid noise and x are features.

- ✔ True

False

Hide Feedback

Because y is still linear in parameter w and we can still learn linear regression from it.

Expectation Maximization

Question 7 — 1 / 1 point

Consider the training data set below. There is some initialization of parameters for which the two clustering methods assign data points to 2 clusters: K-means and Gaussian Mixture Model (GMM) using EM. Answer **True/False** - The two clustering algorithms would produce exactly the same cluster centers (K-means) or cluster means (GMM).

True

✓ False

Hide Feedback

The two clustering would reasonably find the clusters just well. The main difference is that k-means is a hard clustering algorithm which assigns cluster centers as the average of data points lying within that cluster. Whereas, GMM has cluster mean as the weighted average of all the points. GMM assigns a non-zero probability (maybe very small) of each point being into different clusters. GMM assigns soft clustering membership. Therefore, with GMM the mean of the left cluster would be skewed a little bit towards the right and for the right cluster it would be skewed towards the left side. The idea of the question was to highlight the difference between hard and soft membership assignments of the different clustering methods.

Question 8 — 1 / 1 point

Which of the following is true about the application of EM to Gaussian mixture modelling, select **all** correct answers:

✓ **The cost function is negative of the log marginal likelihood**

- ✓ ☑ There is some initialization of model parameters for which EM finds the optimal solution
- ✓ ☑ The marginal likelihood of the data increases in each iteration of EM
- ✓ ☐ For D=2 and K=3, the model has 20 parameters. Here D is dimension of features and K is number of clusters.

▼ Hide Feedback

Reasoning for the final item: the number of model parameters is 2+3*(3+2)=17, where 2 parameters are used for the mixture weights, 3 parameters for each covariance matrix (because it is symmetric) and 2 parameters for each mean.

PCA

Question 9 **1 / 1 point**

In the Principal Component Analysis (select **all** that are **true**):

- ✓ ☑ The principal directions are the directions in the features space along which the data vary the most.
- ✓ ☐ The principal components provide the low-dimensional linear surfaces that are farthest to the observation.
- ✗ ☐ The first principal component optimizes

$$\max_{q_{11}\cdots q_{1D}} \left(\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j=1}^{D} q_{1j}\cdot x_{ij}\right)\right)$$

, where,

$$x_1 = (x_{11}\ldots x_{1D})^T$$

represents the vector of observation 1, similarly

$$q_1 = (q_{11}\ldots q_{1D})^T$$

represents the first principal vector and the column means of N×D data set X are zero.

▼ Hide Feedback

Feedback:

2. The principal components provide the low-dimensional linear surfaces that are closest to the observation. The first D' principal components of the dataset spans the D'-dimensional hyperplane that is closest to the n given observations.

3. The first principal component maximizes the variance of z1 =X.q1 where each of the variables in X has been centered to have zero mean.

Attempt Score: 8 / 9 - 88.89 %

Overall Grade (highest attempt)**:** 8 / 9 - 88.89 %

Done