Advance Machine learning Spring-2021 Homework 01

Your Name

January 20, 2021

1 First Question

Solution : Probability is associated with finding the chance of occurrence of an event when the sample distribution of data is given. Likelihood on the other hand is associated with finding the best distribution of the data given a particular event in the data.

Example : Say we have data of heights of people with mean = 130 and S.D = 5.6 then

Probability = P(X >150 — mean = 130, S.D = 5.6)
Likelihood = L(mean = 130, S.D = 5.6) = P(mean = 130, S.D = 5.6 $|X > 150$)

2 Second Question

List of commonly used evaluation and performance metrics are

- Accuracy metric
- Precision scores
- Recall scores
- F1 score
- Confusion Matrix
- MSE, RMSE, SSE
- Log loss
- ROC AND AUC curves
- Gini coefficent
- Silhouette coefficients, rand index, jaccard index for clustering algorithms
- K-fold cross validation error (not exactly a metric but a technique to evaluate the generalization performance of model)

3 Third question

In a machine learning cycle, once a model has been trained, the model performance indicates how good are the output/predictions given by the trained model on a input data set. Accuracy is one such metric that is used to evaluate a model's performance.

Accuracy : Accuracy is defined as the fraction of correct predictions divided by the total number of predictions, i.e, percentage of correct predictions.

Now there are two scenarios, one where we have a balanced dataset and another where we have an imbalanced dataset.

When data set is balanced, we can use accuracy performance metric given that the cost of mis-classifying true as false and false as true is same, in other words, symmetric loss.

When dataset is imbalanced we cannot use accuracy metric because say for a dataset if the default rate is 10% then even if everything is predicted as 0, the accuracy will still be 90%. Therefore when the dataset is imbalanced we use performance metric like confusion matrix or precision, recall or ROC-AUC curves etc.

Now say we have are building a model to predict weather a patient has a disease or not. Here the cost of a patient actually being positive (meaning having disease) and model predicting the opposite,i.e, negative, is greater than patient actually being negative and model predicting positive. Therefore in such a case where the cost of false negative is high we use performance metrics like recall

Recall = True positive/ true positive + false negative

Similarly when the cost of false positive is high, we use precision as a performance metric

precision = true positive/ true positive + false positive
In both the above cases we use F1 scores and confusion matrix as well

Therefore, in conclusion, for a regression task, we can use performance metric like SSE, MSE, RMSE etc, While for a classification task we commanly use metrics like Confusion matrix, F1, Precision, recall etc.This is a high level view, depending on the problems disccused above weuse suitable metrics accordingly.

4 Fourth question

Model with high bias and low variance suffers from the problem of under fitting meaning model is too simple to approximate the true function. Model with high variance and low bias suffers from problem of over fitting meaning model is too complex in other words ,model has memorized or remembered that data, and hence performs poorly on unseen data. Therefore, both having high bias low variance or high variance low bias is bad for

our model as it performs poorly because of under fitting and over fitting.Hence for the model to Perform well, there needs to be a balance/ trade off between bias and variance. Bias-variance trade off is important to understand how well the model performs on both training and testing/validation data.

5 Fifth question

Regularisation, in general, is a technique using which models are tuned to a preferred level of complexity in order to avoid the problem of overfitting.

L1 regularisation is a technique where we add L1 norm of the weights of the model to the error term of the loss function, i.e,
Loss with L1 regularisation = Error $+\lambda * \sum_{i=1}^{n} |w_i|$

L2 regularisation is a technique where we add squared L2 norm of the weights to the error term of the loss function.
Loss with L2 regularisation = Error $+\lambda * \sum_{i=1}^{n} w_i^2$

Now let us define our model as follows:

$\hat{y} = wx + b$
Let L1 denote loss function with L1 regularisation, i,e,
L1 = $((wx + b)^2 - y) + \lambda * \sum_{i=1}^{n} |w_i|$

Let L2 denote loss function with L2 regularisation, i.e,

L2 = $((wx + b)^2 - y) + \lambda * \sum_{i=1}^{n} w_i^2$

Now gradient descent with L1 and L2 loss function gives :

Gradient descent with L1:
$W_{new} = w - lr * \frac{\partial L1}{\partial w} =$

$w - lr * (2x * (wx + b - y) + \lambda)$ if $w > 0$ else
$w - lr * (2x * (wx + b - y) - \lambda)$ if $w < 0$

Gradient Descent with L2:
$W_{new} = w - lr * \frac{\partial L2}{\partial w} =$
$w - lr * (2x * (wx + b - y) - 2\lambda * w)$
Now let us assume that our model overfits, then we can observe the following:

- in case of L1 regularisation, if w is positive, then the regularisation parameter in L1, i.e, λ will force w to be less positive by subtracting λ from w. if w is negative then λ will force it to be less negative by adding λ to w. Hence L1 regularisation has an effect of pushing W towards zero.

- unlike L1 regularisation, in L2 regularisatin, it depends on both the sign of λ and magnitutde of w as wel (because of $2(\lambda)w$
  In case of L2 regularisation, the equation indicates that here L2 regularisation makes weights decay towards zero but it does not have the effect of pushing the weights towards zero like L1 and Hence L1 is more likely to have zero coefficients than l2.
  This is how we use L1 and L2 regularisation to reduce the over fitting of model.

6 Sixth question

The number of linear regions of the functions that can be computed by a given model is a measure of the model's flexibility In a rectified linear network, to find the number of linear regions, we need to know the distinct number of activation patterns in the network because from [1] we can see that the distinct number of linear regions in a network corresponds to the distinct number of activation patterns in the network.

For example, say there are 5 activation patterns for a given choice of parameters of the rectifier network then the upper bound for the number of maximal linear regions is 5.

From [2] we see that a network with N relu's units have at max 2^n activation patterns. This is a loose bound as not all activation patterns are active at all time. Many activation units are empty for a given input. Hence, this is a loose bund. Since there are 2^N max activation patterns for a given parameters of the model, there can at most be 2^N linear regions

Referred papers :

[1] : `https://www.researchgate.net/publication/322539221_Notes_on_the_number_of_linear_regions_of_deep_neural_networks`

[2] : `https://papers.nips.cc/paper/2014/file/109d2dd3608f669ca17920c511c2a41e-Paper.pdf`