CSE 4/574

Introduction to Machine Learning
Homework 1
Total Marks 50

Your Name:
Your email ID:
Your UB Person ID:

1. [5] Show that if two variables x and y are independent, then their covariance is zero.

Formula for covariance for 2 variables x and y =

cov[x,y] = E [(x -E[x])(y -E[y])]

(opening the brackets & multiplying & using the fact that for independent x,y E[E[x]] = E[x])
cov[x,y] = E[xy] - E[x]E[y] - E[x]E[y] + E[x]E[y]

cov[x,y] = E[xy] - E[x]E[y]

As for independent variables, E[xy] = E[x]E[y],
And E[A] is mean of 'A'

cov[x,y] = E[x]E[y] - E[x]E[y] = 0

Hence, proved that covariance for 2 independent variables (x,y) is 0.

2. [5] Using the definition $var(f) = E[(f(x) - E[f(x)])^2]$ show that var[f(x)] satisfies $var(f) = E[f(x)^2 - E[f(x)]^2]$.

As given in question, formula for variance for x = var(f) = E[(f(x)–E[f(x)])2]

when we try to open & manipulate the given formula using identity (a-b)^2 = a^2 -2ab + b^2 ,
var(f) = E[f(x)^2 -2*f(x)*E[x] + E[f(x)]^2]

As E[E[x]] = E[x],

var(f) = E[f(x)^2] -2*E[f(x)]*E[f(x)] + E[f(x)^2] = E[f(x)^2] - E[f(x)]^2 = E[f(x)^2 - E[f(x)]^2]
Hence proved as stated in the question.

3. [5] With a 0.5% chance of selecting a ‘faulty’ coin with both heads, and a 99.5% chance of picking up a ‘fair’ coin, you are running your experiments of coin flipping and you end up getting 10 consecutive heads. What’s the probability that you picked up the fair coin, given the information above? Hint: Use Bayes Formula

Given P(Fair_coin) = 995/1000   & P(Faulty_coin) = 5/1000
For 10 consecutive heads-> P(Heads | Fair_coin) = (1/2)^10
For 10 consecutive heads-> P(Heads | Faulty_coin) = 1 (as faulty coin has no tails)

Thus, For 10 consecutive heads-> P(Heads)
= P(Faulty_coin)*P(Heads | Faulty_coin) + P(Fair_coin)P(Heads | Fair_coin)
= 5/1000*1 + 995/1000*(1/2)^10 = 0.00597167968
Using Bayes Formula:
For 10 consecutive heads->
P(Fair_coin | Heads) = [P(Fair_coin) X P(Heads | Fair_coin)] / P(Heads)
= [995/1000 * (1/2)^10] / 0.00597167968

P(Fair_coin | Heads) = 0.16271463508

4. [10] Explain Univariate and multivariate Normal distribution.

The main difference between univariate and multivariate distribution is that univariate considers only one variable for distribution & multivariate considers more than 2 variables for distribution.

In terms of Normal distribution, univariate considers the effect of one variables on the output. On other hand, Multivariate considers the combined effect of all random variables on the output.

In univariate normal distribution, variance is calculated to gauge uncertainty in the distribution. On other hand, Multivariate normal distribution uses covariance of random variables in the question to gauge uncertainty in distribution.Similarly even the dimensions of both univariate & multivariate normal distributions are different.

5. [5] What is the difference between Supervised Learning and Unsupervised Learning?

The whole idea behind machine learning is to learn the distribution (ground truth) governing the source of the data. This learning can happen via variety of ways. Two most important ways are Supervised learning & unsupervised learning. In supervised learning we are given a labeled data, using which we try to learn ground truth function using various method like SGD,SVM, etc. However in Unsupervised learning, given data is not labeled. Hence for Unsupervised learning, we tend to use different methods like clustering, k-means, EM, SVD. This methods rely on given data's embedded statistics in place of label.

6. [10] What is Regularization and what kind of problems does regularization solve? Describe Ridge Regression.

As mentioned in one of the previous answers that we aim to learn the distribution (ground truth) governing the source of the data rather than learning the data itself. Hence we try to improve the generalisation of the learned model (i.e. ability to mimic the ground truth) & try avoid the overfitting of the learned model (i.e. tendency to memorise the sample data). Overfitting usually happens when the learned model is too complex relative to the ground truth function. To avoid overfitting we try to use regularization, which is to reduce the complexity of the learned model. Regularization can be implemented in may ways (directly or indirectly.). However one of the most common ways to implement regularization is introducing an another term for regularization in the cost function. This causes the suppression of weight parameters of learned models in variety of ways. The kind of suppression depends upon the nature of regularization term used.

For example, one kind of regularization is called Ridge regularization. Ridge regularization is also called L2 regularization. Ridge reg. introduces the sum of squares of the parameter weights to the cost function. This causes the parameters weights to suppress in sense that learned model is not too much dependent upon few parameter with large weights. Thus Ridge reg. tries to distribute the weights more equally rather than go for an alternative sparse solution (as in Lasso Regularization). Graphically Ridge Reg. mimics a circle.

7. [5] What is Overfitting? What are the different methods that we have discussed to resolve overfitting?

As mentioned in one of the previous answers that we aim to learn the distribution (ground truth) governing the source of the data rather than learning the data itself. Hence we try to improve the generalisation of the learned model (i.e. ability to mimic the ground truth) & try avoid the overfitting of the learned model (i.e. tendency to memorise the sample data). Overfitting usually happens when the learned model is too complex relative to the ground truth function. To avoid overfitting, we can use a number of ways. Most important of them is regularization, which is to reduce the complexity of the learned model.

Regularization can be implemented in may ways (directly or indirectly.). However one of the most common ways to implement regularization is introducing an another term for regularization in the cost function. This causes the suppression of weight parameters of learned models in variety of ways. The kind of suppression depends upon the nature of regularization term used.

Secondly, we can limit the number of loops, iterations, or learning of data so that we can achieve the best Test error instead of just focussing of Training error. We can also introduce cross-validation testing to get a better idea about continuous improvement in learning rate.

Thirdly, the best thing is to get more data or more diverse data than you already have.

Lastly, we can also use indirect methods like dropout layers to reduce overfitting.

8. [5] Given a sample collection you fit a degree $M = 4$ polynomial and found that both training and testing error are "0". Now you plan to explore a bit more and change the value for M slightly up to 6 and down to 2. What would you expect in terms of the resulting curves fitting the data and their performance in the test collection?

In the above example, we already have achieved best possible training & testing error of "0" with given complexity of degree (M =4).

Thus, increase in complexity of degree (M =6) will cause overfitting of the data. This will result in training error of "0" but will increase the testing 'error to greater than "0".

On other hand, decreasing complexity of degree (M =2), will might cause under fitting of the data i.e. the model is too simple to learn the complex ground truth. Hence, it will result in increase in testing error to greater than "0".