Mustapha Tidoo Yussif

3rd April 3, 2019

Homework 3

Question 1.

The independent identically distributed (IID) assumption is used to explain some relationship between random variables in probability theory. It states that all data points are independent of each other although each data point comes from the same probability distribution. Flipping a coin is a good example to explain IID. Every flip uses the same coin, but the second flip of the coin does not depend on the first flip. The importance of the IID assumption includes (but not limited to):

1. IID assumption makes it possible for us to compute the joint probability of the data by simply taking the product of the probabilities of all the individual data points. For example, if we want to calculate the joint probability P (x1, x2, x3, x4… xn), we will multiply the individual probabilities as: P(x1, x2, x3, x4 …xn) = p(x1) * p(x2) …p(xn).
2. Also, it simplifies the processing of parameter estimation via maximum likelihood estimation.

Question 2.

I will choose the maximum likelihood regression with Gaussian mean noise. For the most part, natural data have inherent noise. This predicts vanilla linear regression imperfect because it completely ignores the noise errors in the model. On the contrary, Maximum Likelihood regression with Gaussian mean noise assumes normality of the noise with a mean of 0 and some unknown variance. Gaussian seems to be the good choice because when you draw a line to pass through linearly plotted data points, the errors are symmetric about were the line drawn. Unlike the vanilla linear regression, the "*y*" value is estimated using three parameters namely the slope, the intercept and the variance of the noise distribution. Because the ML regression includes the noise error term, it has a minimal residual error.

Question 3.

Though both conventions provide good and consistent estimations for the true variance of the distribution when the dataset is relatively large, 1/m version, which is the maximum likelihood estimation of the variance turns to be biased when the dataset is relatively small, and the octave version turns to be unbiased. The disparity only affects the accuracy of the model if the dataset is relatively small or moderate. But when the dataset is large, the difference becomes negligible and does not affect the model accuracy.

Question 4.

(a) If the features of each example are independent, then the mean for the multivariate model will be $\mu = [\mu_1, \mu_2, \mu_3 \ldots \mu_n]$ and the covariate is $\sum = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$ which is essentially the diagonal values (the same as variances in the vanilla gaussian distribution) since the rest of the elements in the matrix are zeros. Therefore, the model takes n means and n variances.

(b) . If we ignored the independence assumption in modeling the distribution, the mean will still be $\mu = [\mu_1, \mu_2, \mu_3 \ldots \mu_n]$, but the covariance will change, $\sum = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_n^2 \\ \vdots & \ddots & \vdots \\ \sigma_n^2 & \cdots & \sigma_n^2 \end{bmatrix}$. Hence the model takes n means and n X n variances.

Question 5.

1. Cross-validation is often used when the dataset is small and not much to be split into training, validation and test datasets. This is not the case in deep learning since deep learning requires a lot of data (to typically learn a complex model) which is enough and can be held out to estimate performance reliably.
2. Also, cross-validation is not used in deep learning because it is computationally expensive.

Question 6

Parameters are the properties of the training data that are learned during the training of the model. The objective during training is to find or learn the features of the data that optimizes some loss function. In logistic regression, the weights and the biases are the parameters of the model. **Hyperparameters** on the hand are the values that must be decided before you start training your model. For example. In k-means clustering, k which is the number of clusters expected is a hyperparameter. Also, in logistic regression, lambda (the regularization value) is a hyperparameter.

Question 7.

The datasets for anomaly detection problems have large discrepancies in terms of positive and negative examples. For the most part, positive examples are very small. Contrary to that, the negative examples are many. This makes it suitable to use anomaly detection algorithms to learn the data. Using Gaussian distribution for anomaly detection is convenient. It assumes normality of the training data. Assuming normality of the data simplifies the estimation of μ and σ^2 via maximum likelihood estimation:

$$\mu_i = \frac{1}{N}\sum_{k=1}^{N} x_i^{(k)}$$

$$\sigma_i^2 = \frac{1}{N}\sum_{k=1}^{N} (x_i^{(k)} - \mu_i)^2$$

And thus, makes it possible to compute $p(x) = \prod_i p(x_i; \mu_i, \sigma_i^2)$. Which is used to flag *x_test* as an anomaly if *P(x_test)* is less than the specified threshold and non anomaly if its greater than the threshold. Aside from making the parameter estimation easy, the normality assumption used in Gaussian distribution is often a good approximation of reality making Gaussian distribution a good model for anomaly detection.

Nevertheless, Gaussian distribution is not a good choice for anomaly detection when the data is not normally distributed.

Question 8.

One of the ultimate aims of machine learning is to avoid overfitting. As a result, we normally choose a model that is moderately complex to capture the necessary features in the data. Often time some of these models turn to overfit if not checked. This is when regularization comes to play. Regularization discourages learning overly complicated function to avoid the risk of overfitting. This is done by controlling the regularization parameter (Greek lambda). The more complex the model, the larger the penalty, which means the value of lambda is tuned to force the coefficients of the model terms to be closer to zero. Choosing the regularization parameter value is tricky and techniques such as cross-validation is often used to determine the best value. On the other hand, if the regularization parameter is infinitely very small, the complexity of the model is not reduced.

The larger the term’s coefficient size, the larger the penalty, which basically means the more the tuning parameter forces the coefficient to be closer to zero. Choosing the value to use for the tuning parameter is critical and can be done using a technique such as cross-validation.

Question 9.

Maximum Likelihood for regression and Maximum A Posteriori (MAP) regression, is both a method for estimating the best parameters for linear regression models. They are similar in the sense that both compute single estimate (say the mean of the distribution) instead of full delivery. Some of the difference between the ML and the MAP regression is that MAP regression does not require regularization term since MAP regression models do not overfit. ML regression, on the other hand, needs a regularization term to avoid the risk of overfitting. Also, the independence assumption is made to use ML regression to estimate the model parameters which is not the case in MAP regression.

Question 10.

Using just the prediction errors of a model to evaluate the model is not a good practice. This is because it can lead to overestimation of the goodness of a model. For example, if the prediction errors of a model are -3, -2, -1, 0, 1, 2, and 3, we might be tempted to say the model is a perfect one since it has zero prediction errors (3 + 2+ 1+ 0 + -1 + -2+ -3 = 0) over the dataset. To curb this, Root Mean Squared Error is used. How does it solve this challenge? At a point, it squares all the errors and averages all the values (Mean Squared Error). Squaring the errors makes it impossible to get zero Mean Squared Error. Finally, we take the square root of the Mean Squared Error to make the scale of the errors to be the same as the scale of targets.