**MACHINE LEARNING -** **WORKSHEET 1**

**1. Which of the following methods do we use to find the best fit line for data in Linear Regression?**

A) Least Square Error

B) Maximum Likelihood

C) Logarithmic Loss

D) Both A and B

**ANSWER---A) LEAST SQUARE ERROR**

**2. Which of the following statement is true about outliers in linear regression?**

A) Linear regression is sensitive to outliers

B) linear regression is not sensitive to outliers

C) Can’t say

D) none of these

**ANSWER---A) LINEAR REGRESSION IS SENSITIVE TO OUTLINERS**

**3. A line falls from left to right if a slope is \_\_\_\_\_\_?**

A) Positive

B) Negative

C) Zero

D) Undefined

**ANSWER---B) NEGATIVE**

**4. Which of the following will have symmetric relation between dependent variable and independent variable?**

A) Regression

B) Correlation

C) Both of them

D) None of these

**ANSWER---B) COLLERATION**

**5. Which of the following is the reason for over fitting condition?**

A) High bias and high variance

B) Low bias and low variance

C) Low bias and high variance

D) none of these

**ANSWER---C) LOW BIAS AND HIGH VARIANCE**

**6. If output involves label then that model is called as:**

A) Descriptive model

B) Predictive modal

C) Reinforcement learning

D) All of the above

**ANSWER---B) PREDICTIVE**

**7. Lasso and Ridge regression techniques belong to \_\_\_\_\_\_\_\_\_?**

A) Cross validation

B) Removing outliers

C) SMOTE

D) Regularization

**ANSWER---D) REGULARIZATION**

**8. To overcome with imbalance dataset which technique can be used?**

A) Cross validation

B) Regularization

C) Kernel

D) SMOTE

**ANSWER---D) SMOTE**

**9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses \_\_\_\_\_ to make graph?**

A) TPR and FPR

B) Sensitivity and precision

C) Sensitivity and Specificity

D) Recall and precision

**ANSWER---A) RECCALL AND PRECISION**

**10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.**

A) True B) False

**ANSWER---B) FALSE**

1**1. Pick the feature extraction from below:**

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

C) Removing stop words

D) Forward selection

**ANSWER---C) REMOVING STOP WORDS**

**12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?**

A) We don’t have to choose the learning rate.

B) It becomes slow when number of features is very large.

C) We need to iterate.

D) It does not make use of dependent variable.

**ANSWER---A) WE DON’T HAVE TO CHOOSE THE LEARNING RATE**

**B) IT BECOMES SLOW when number of features is very large.**

**C) We need to iterate.**

**13. Explain the term regularization?**

**ANSWER---** The term ‘regularization’ refers to a set of techniques that regularizes learning from particular features for traditional algorithms or neurons in the case of neural network algorithms .It normalizes and moderate's weights attached to a feature or a neuron so that algorithms do not rely on just a few features or neurons to predict the result. This technique helps to avoid the problem of overfitting.

To understand regularization, let’s consider a simple case of linear regression. Mathematically, linear regression is stated as below:

y = w0 + w1x1 + w2x2 + ….. + wnxn

where y is the value to be predicted;

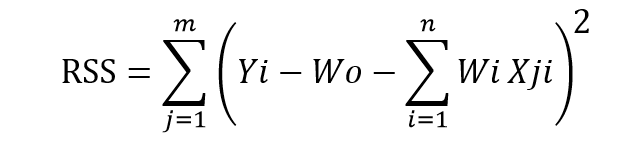
x1, x2, …., xn are features that decides the value of y;

w0 is the bias;

w1, w2, ….., wn are the weights attached to x1, x2, …., xn relatively.

Now to build a model that accurately predicts the y value, we need to optimize above mentioned bias and weights.

To do so, we need to use a loss function and find optimized parameters using gradient descent algorithms and its variants .The loss function called ‘the residual sum of square’ is mostly used for linear regression. Here’s what it looks like :



bias (or intercept) and weights (also identified as parameters and coefficients) using the optimization algorithm (gradient descent) and data. If your dataset does have noise in it, it will face overfitting problem and learned parameters will not generalize well on unseen data .To avoid this, you will need to regularize or normalize your weights for better learning.

There are three main regularization techniques, namely:

Ridge Regression (L2 Norm)

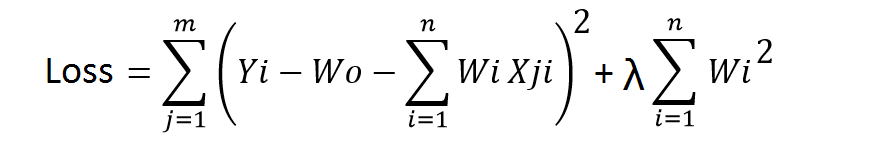
Lasso (L1 Norm)

Dropout

Ridge and Lasso can be used for any algorithms involving weight parameters, including neural nets. Dropout is primarily used in any kind of neural networks e.g. ANN, DNN, CNN or RNN to moderate the learning.

**14. Which particular algorithms are used for regularization?**

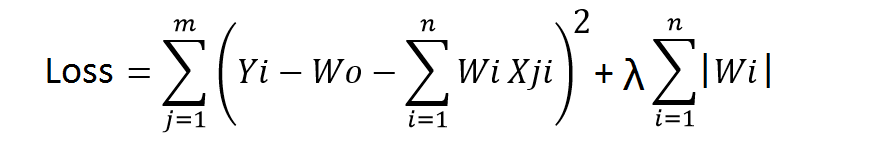
**ANSWER---** Ridge Regression (L2 Regularization) Ridge regression is also called L2 norm or regularization .When using this technique, we add the sum of weight’s square to a loss function and thus create a new loss function which is denoted thus:



As seen above, the original loss function is modified by adding normalized weights. Here normalized weights are in the form of squares .You may have noticed parameters λ along with normalized weights. λ is the parameter that needs to be tuned using a cross-validation dataset. When you use λ=0, it returns the residual sum of square as loss function which you chose initially. For a very high value of λ, loss will ignore core loss function and minimize weight’s square and will end up taking the parameters’ value as zero .Now the parameters are learned using a modified loss function. To minimize the above function, parameters need to be as small as possible. Thus, L2 norm prevents weights from rising too high

Lasso Regression (L1 Regularization)

Also called lasso regression and denoted as below:

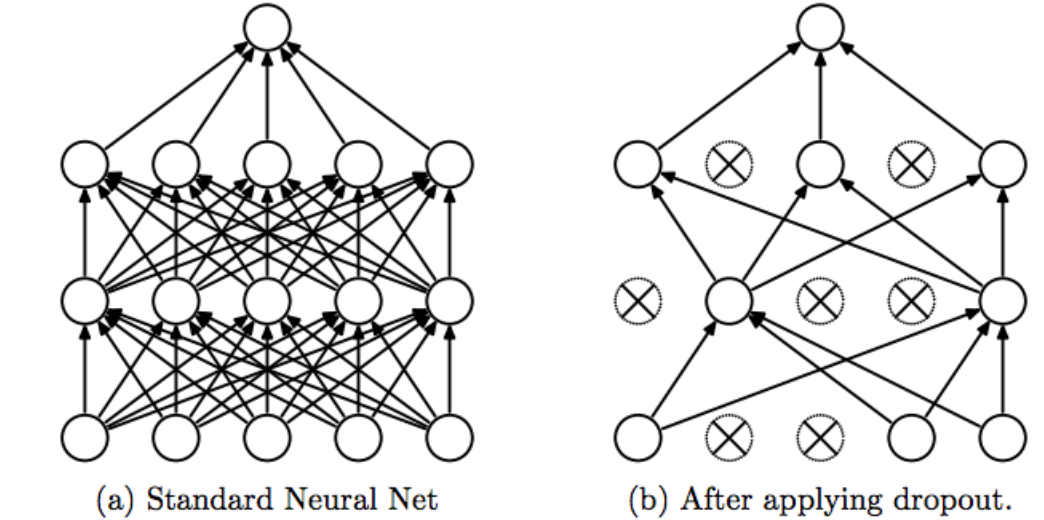


This technique is different from ridge regression as it uses absolute weight values for normalization. λ is again a tuning parameter and behaves in the same as it does when using ridge regression .As loss function only considers absolute weights, optimization algorithms penalize higher weight values.

In ridge regression, loss function along with the optimization algorithm brings parameters near to zero but not actually zero, while lasso eliminates less important features and sets respective weight values to zero. Thus, lasso also performs feature selection along with regularization.

Dropout

Dropout is a regularization technique used in neural networks. It prevents complex co-adaptations from other neurons .In neural nets, fully connected layers are more prone to overfit on training data. Using dropout, you can drop connections with 1-p probability for each of the specified layers. Where p is called keep probability parameter and which needs to be tuned.



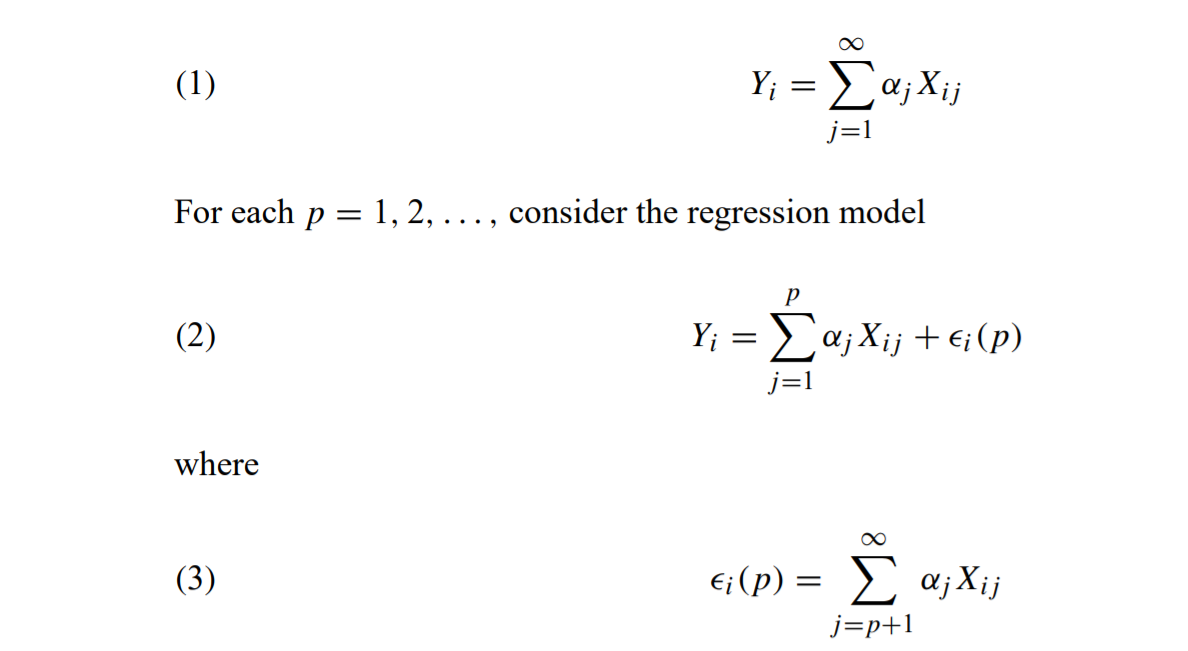
With dropout, you are left with a reduced network as dropped out neurons are left out during that training iteration .Dropout decreases overfitting by avoiding training all the neurons on the complete training data in one go. It also improves training speed and learns more robust internal functions that generalize better on unseen data. However, it is important to note that Dropout takes more epochs to train compared to training without Dropout (If you have 10000 observations in your training data, then using 10000 examples for training is considered as 1 epoch).

Along with Dropout, neural networks can be regularized also using L1 and L2 norms .For real-world applications, it is a must that a model performs well on unseen data. The techniques we discussed can help you make your model learn rather than just memorize.

**15. Explain the term error present in linear regression equation?**

**ANSWER---** It is often said that the error term in a regression equation represents the effect of the variables .that were omitted from the equation. This is unsatisfactory, even in simple contexts, as the following discussion should indicate. Suppose subjects are IID, and all variables are jointly normal with expectation 0. Suppose the explanatory variables have variance 1. The explanatory variables may be correlated amongst themselves, but any p of them have a non-singular p-dimensional distribution.

The parameters αj are real. Let



The αj are identifiable. If the Xij are independent for j = 1, 2,..., the standard assumptions hold, and i(p) does indeed represent the effect on Yi of the omitted variables{Xij : j = p+1,...},at least in an algebraic sense. On the other hand, if the Xij are dependent, the matter is problematic .If we take (1–3) as written, then i(p)represents the effect on Yi of the omitted variables—but i(p) is correlated with the explanatory variables. The standard assumptions fail, and fitting (2) to data for i = 1,...,n will estimate the wrong parameters. If i(p) is replaced by i(p)⊥, namely, the part of i(p) independent of Xi1,...,Xip, we have a bona fide regression model, but with different α’s. There is no easy way out of the difficulty. The conventional interpretation for error terms needs to be reconsidered. At a minimum, something like this would need to be said: the error term represents the combined effect of the omitted variables, assuming that

(i) the combined effect of the omitted variables is independent of each variable included in

the equation,

(ii) the combined effect of the omitted variables is independent across subjects,

(iii) the combined effect of the omitted variables has expectation 0.