**MACHINE LEARNING – WORKSHEET 3**

1. Give short description each of Linear, RBF, Polynomial kernels used in SVM

Answer: The linear, polynomial and RBF are simply different in case of making the hyperplane decision boundary between the classes

The Linear kernel is the simplest kernel function. It is given by the inner product <x,y> plus an optional constant**c**. Kernel algorithms using a linear kernel are often equivalent to their non-kernel counterparts. It is less time consuming function.

k(x, y) = x^T y + c

The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data is normalized.

k(x, y) = (alpha x^T y + c)^d 

Adjustable parameters are the slope alpha, the constant term c and the polynomial degree d.

Apart from the classic linear kernel which assumes that the different classes are separated by a straight line, a RBF (radial basis function) kernel is used when the boundaries are hypothesized to be curve-shaped.

k(x, y) = expleft(- gamma lVert x-y rVert ^2 ) 

The adjustable parameter sigma plays a major role in the performance of the kernel, and should be carefully tuned to the problem at hand. If overestimated, the exponential will behave almost linearly and the higher-dimensional projection will start to lose its non-linear power. In the other hand, if underestimated, the function will lack regularization and the decision boundary will be highly sensitive to noise in training data.

2. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit of model in regression and why??

Answer: R square is the better measure for linear regression, R2 represents the proportion of the variance in our data which is explained by your model; the closer to one, the better the fit.

The residual sum of squares (RSS) is the sum of the squared distances between your actual versus your predicted values:

RSS=∑i=1n(yi−y^i)2

Where yi is a given datapoint and y^i is our fitted value for yi.

The actual value of RSS get depends largely on the **scale** of our response variable. Taken alone, the RSS isn't so informative.

3. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

Answer: The sum of squares total, denoted SST, is the squared differences between the observed *dependent variable* and its mean.

TSS = SSE + SSR

ESS: It is the sum of squares due to regression, or SSR or ESS. It is the sum of the differences between the *predicted* value and the mean of the *dependent variable*. Think of it as a measure that describes how well our line fits the [data](https://365datascience.com/numerical-categorical-data/).

SSR = ∑i =1 (yˆi − y¯) 2

SSE: It is a residual sum of square given by

SSE = ∑i =1 (yi − yˆi) 2

It is the difference between the square of actual value minus predicted value. Also known as RSS(Residual Sum of Squares)

4. What is Gini –impurity index?

Answer: The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits. Gini Impurity tells us what is the probability of misclassifying an observation lower the Gini the better the split. In other words the lower the likelihood of misclassification.

5. Are unregularized decision-trees prone to overfitting? If yes, why?

Answer: Yes, unregularized decision tree prone to overfit, overfitting can be one problem that describes if your model no longer generalizes well. Overfitting happens when any learning processing overly optimizes training set error at the cost test error. While it’s possible for training and testing to perform equality well in cross validation, it could be as the result of the data being very close in characteristics, which may not be a huge problem. In the case of decision tree’s they can learn a training set to a point of high granularity that makes them easily overfit. Allowing a decision tree to split to a granular degree, is the behavior of this model that makes it prone to learning every point extremely well — to the point of perfect classification — ie: overfitting.

Following steps can we use to avoid overfitting:

* Use a test set that is not exactly like the training set, or different enough that error rates are going to be easy to see.
* Ensure to have enough data.
* Reduce the complexity of the decision tree model
* Use decision trees in an ensemble
* Reduce the dimensionality of your data

6. What is an ensemble technique in machine learning?

Answer: Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model. It create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. It uses the voting, to come to conclusion.

7. What is the difference between Bagging and Boosting techniques?

Answer:

|  |  |
| --- | --- |
| **Bagging** | **Boosting** |
| Various training data subsets are randomly drawn with replacement from the whole training dataset. | Each new subset contains the components that were misclassified by previous models. |
| Bagging attempts to tackle the over-fitting issue. | Boosting tries to reduce bias. |
| If the classifier is unstable (high variance), then we need to apply bagging. | If the classifier is steady and straightforward (high bias), then we need to apply boosting. |
| Every model receives an equal weight. | Models are weighted by their performance. |
| Objective to decrease variance, not bias. | Objective to decrease bias, not variance. |
| It is the easiest way of connecting predictions that belong to the same type. | It is a way of connecting predictions that belong to the different types. |
| Every model is constructed independently. | New models are affected by the performance of the previously developed model. |

8. What is out-of-bag error in random forests?

Answer: Out of bag (OOB) score is a way of validating the Random forest model. Out-of-bag error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating to sub-sample data samples used for training.

9. What is K-fold cross-validation?

Answer: Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation. When a specific value for k is chosen, it may be used in place of k in the reference to the model, such as k=10 becoming 10-fold cross-validation.

10. What is hyper parameter tuning in machine learning and why it is done?

Answer: A hyperparameter is a parameter whose value is set before the learning process begins. Some examples of hyperparameters include penalty in logistic regression and loss in stochastic gradient descent.

Method of optimizing

**Grid Search**

In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

**Random Search**

RandomizedSearchCV solves the drawbacks of GridSearchCV, as it goes through only a fixed number of hyperparameter settings. It moves within the grid in random fashion to find the best set hyperparameters. This approach reduces unnecessary computation.

11. What issues can occur if we have a large learning rate in Gradient Descent?

Answer: When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. When the learning rate is too small, training is not only slower, but may become permanently stuck with a high training error. When using high learning rates, it is possible to encounter a positive feedback loop in which large weights induce large gradients which then induce a large update to the weights. If these updates consistently increase the size of the weights, then [the weights] rapidly moves away from the origin until numerical overflow occurs.

12. What is bias-variance trade off in machine learning?

Answer: In our machine learning model, if model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it’s going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can’t be more complex and less complex at the same time.

13. What is the need of regularization in machine learning?

Answer: It is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, *this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.*

A simple relation for linear regression looks like this. Here Y represents the learned relation and *β represents the coefficient estimates for different variables or predictors(X).*

*Y ≈ β0 + β1X1 + β2X2 + …+ βpXp*

It will adjust the coefficients based on your training data. *If there is noise in the training data, then the estimated coefficients won’t generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.*

*Lasso and Ridge are regularization techniques.*

14. Differentiate between Adaboost and Gradient Boosting

|  |  |
| --- | --- |
| Gradient boosting | Adaptive Boosting |
| This approach trains learners based upon minimising the loss function of a learner (i.e., training on the residuals of the model) | This method focuses on training upon misclassified observations. Alters the distribution of the training dataset to increase weights on sample observations that are difficult to classify. |
| Weak learners are decision trees constructed in a greedy manner with split points based on purity scores (i.e., Gini, minimise loss). Thus, larger trees can be used with around 4 to 8 levels. Learners should still remain weak and so they should be constrained (i.e., the maximum number of layers, nodes, splits, leaf nodes) | The weak learners incase of adaptive boosting are a very basic form of decision tree known as stumps. |
| All the learners have equal weights in the case of gradient boosting. The weight is usually set as the learning rate which is small in magnitude. | The final prediction is based on a majority vote of the weak learners’ predictions weighted by their individual accuracy. |

15. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

Answer: Logistic regression is known and used as a linear classifier. It is used to come up with a hyper*plane* in feature space to separate observations that belong to a class from all the other observations that do *not* belong to that class. The decision boundary is thus *linear.* Logistic regression is an exercise in predicting (regressing to — one can say) discrete outcomes from a continuous and/or categorical set of observations. Each observation is independent and the probability *p* that an observation belongs to the class is some ( & same!) function of the features describing that observation.