Natasha Kubiak

HW 3.1

ECE 517

1) Explain the concepts of Risk and Empirical Risk.

Suppose we have a learning machine such that it is able to construct a set of parametric estimation functions, f(x, α) where α is a set of parameters to be adjusted. Now we can assume that the estimation function is

$$f(x_n, \alpha) = y_n + e_n$$

And we can assume e_n to be the estimation error. Therefore we will want to minimize the the convex function of the error using Risk Minimization.Where we can define the risk function to be

$$R(\alpha) = \int_{xy} L(y, f(x, \alpha)dF(x, y)$$

Risk is then the integral of the Loss function where L is a measure of discrepancy between the output and the label.

We acknowledge however that we cannot solve for $R(\alpha)$ and must take an approximation instead of taking the integral for our risk function. We instead take the summation of our Loss function, such that we are calculating the Empirical risk instead. Where the Empirical Risk is define as

$$R_{emp}(\alpha) = \sum_{n=1}^{N} L(y_n, f(x_n, \alpha))$$

2) Explain the concepts of complexity and overfitting.

Within our Learning Machine Model we need to find the optimal model complexity. This complexity refers to the proportions of our training set such as dimensions, variables, and classifiers, as to not create an over or under fitting of our model.

Overfitting is when our Learning Machine has learnt the given training data too well such that it performs excellently with the training data but will not account for discrepancies. A model which is over-fit will have a low bias and high variance, and it’s performance would vary with data outside of the training set.

3) Introduce the concept of VC dimension.

The Vapnik Chervonenkis Dimension gives us a measure of the complexity of linear functions. Where a linear classifier is the size of the largest set of points that can be shattered as to remain linearly independent in the VD Dimension. We can consider a set of m points in R^n

where we can choose any point to be the origin. With m number of points , we can shatter them with oriented hyperplanes if and only if the position vectors of the remaining points are linearly independent.

4) Enunciate and interpret the VC theorem that describes the bound on the actual risk.

We can take the Empirical Risk to be

$$R_{emp}(\alpha) = \frac{1}{2N}\sum_{n=1}^{N} |y - f(x, \alpha)|$$

We then define that our loss function |y − f(x, α)| can only take the values 0 or 1. Then our bound will be independent of our probability distribution $1 - \eta$ such we can define our actual risk to be

$$R(\alpha) = R_{emp}(\alpha) + \frac{\sqrt{h(log(2N/h)+1)-log(\eta/4)}}{N}$$

Where our Structural Risk Minimization will consist on choosing a dimension h that is small enough so that we may minimize the bound on actual risk.

5) Introduce the SVM criteria

For our Support Vector Machine, we will want to find the hyperplane in a space of as N dimensions as described previously in the VP Dimensions. For the SVM, we will want to minimize the empirical risk plus the structural risk through margin maximization. Thus we can

minimize

$$L_p(w, \xi_n) = \frac{1}{2}||w||^2 + C\sum_{n=1}^{N} \xi_n$$

Subject to

$$y_n(w^T x_n + b) > 1 - \xi_n \quad where\, \xi_n \geq 0$$

Where we find the hyperplane of best fit such that it maximizes the margin such that it has the maximum distance between the data points of our classes.

6) Develop the analysis that leads to the dual solution of the SVM, and its main results.

We will want to minimize the Empirical Risk plus the Structural Risk for our Support Vector Machine.

We minimize

$$L_p(w, \xi_n) = \frac{1}{2}||w||^2 + C\sum_{n=1}^{N} \xi_n$$

Subject to

$$y_n(w^T x_n + b) > 1 - \xi_n \quad where\, \xi_n \geq 0$$

We optimize the machine using the Lagrange minimization where

$$L_{Lagrange} = F(w) - \alpha g(w)$$

Where $\alpha \geq 0$ is our duel variable.

To optimize using the lagrange minimization, we compute the gradient with respect to primal variable w.

$$\Delta_w F(w) - \alpha g(w) = 0$$

Which leads to finding the Karush Kuhn Tucker (KKT) conditions which we can then find the value of the dual variable

7) Describe the properties of the Support Vectors.

A support Vector is data points that will be the closest to our separating hyperplane.The Support Vector is used to maximize the margin distance from our separating hyperplane and touch the margin. Such that these points decide the hyperplane that best fits our training data.