Machine Learning End Term Exam

Ahmad Shayaan IMT2014004
H Vijaya Sharvani IMT2014022
Indu Ilanchezian IMT2014024

December 11, 2017

Question 2

To derive the solution to the modified linear regression leads to the generalized form of ridge regression.

Solution:-

Given the attribute $x_i = \hat{x_i} + \epsilon_i$, where the $\hat{x_i}$ are the true measurements and ϵ_i is the zero mean vector with covariance matrix $\sigma^2 I$
Modified loss function

$$W^* = argmin_w E_\epsilon \sum_{i=1}^{n} (y_i - W^T(\hat{x_i} + \epsilon_i))^2$$

Where W is the transformation vector.

$$W^* = argmin_W E_\epsilon ||Y - (X + \epsilon)W||_2^2 \tag{1}$$

Where

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$X = \begin{bmatrix} \hat{x}_1^T \\ \hat{x}_2^T \\ \vdots \\ \hat{x}_n^T \end{bmatrix}$$

$$\epsilon = \begin{bmatrix} \epsilon_1^T \\ \epsilon_2^T \\ \vdots \\ \epsilon_n^T \end{bmatrix}$$

Expanding right hand side of equation 1.

$$E_\epsilon||Y-(X+\epsilon)W||_2^2 = E_\epsilon\Big[(Y-(X+\epsilon)W)^T(Y-(X+\epsilon)W)\Big]$$

$$= E_\epsilon\Big[Y^TY + W^T(X+\epsilon)^T(X+E) - 2W^T(X+E)^TY\Big] \tag{2}$$

To minimize the equation we will differentiate eq 2 wrt W.

$$\frac{\partial E_\epsilon\Big[Y^TY + W^T(X+\epsilon)^T(X+\epsilon)W - 2W^T(X+E)^TY\Big]}{\partial W} = 0$$

We know that $\frac{\partial E(f(x))}{\partial} = E\frac{\partial f(x)}{\partial x}$.

$$E_\epsilon\Big[\frac{\partial Y^TY}{\partial W} + \frac{\partial W^T(X+\epsilon)^T(X+\epsilon)W}{\partial W} - 2\frac{\partial W^T(X+E)^TY}{\partial W}\Big] = 0$$

$$E\epsilon\big[2(X+\epsilon)^T(X+\epsilon)W - 2(X+\epsilon)^TY\big] = 0$$

$$2E_\epsilon\big[(X+\epsilon)^T(X+\epsilon)W\big] - 2E_\epsilon\big[(X+\epsilon)^TY\big] = 0$$

$$E_\epsilon\big[(X^TX + \epsilon^T\epsilon + 2\epsilon^TX)W\big] = E_\epsilon\big[(X+\epsilon)^TY\big])$$

$$E_\epsilon(X^TXW) + E_\epsilon(\epsilon^T\epsilon W) + 2E_\epsilon(\epsilon^TXW) = E_\epsilon(X^TY) + E_\epsilon(\epsilon^TY)$$

We know that E(AB) = E(A)E(B) if A and B are independent variables and $E_f(h(x)) = \int_{-\infty}^{\infty} h(x)f(x)dx$.

$$\sum_{i=1}^{n} X^TXWP(\epsilon_i) + E_\epsilon(\epsilon\epsilon^T)E_\epsilon(W) + 2E_\epsilon(X)E_\epsilon(\epsilon) = \sum_{i=1}^{n} X^TYP(\epsilon_i) + E_\epsilon(Y)E_\epsilon(\epsilon)$$

We know that the noise is a zero mean Gaussian noise therefore E(ϵ) = 0

$$(X^TX + \sigma^2I)W = X^TY$$

$$W = (X^TX + \sigma^2I)^{-1}X^TY$$

therefore the solution of the minimization is

$$W^* = (X^TX + \sigma^2I)^{-1}X^TY$$

This solution is same as the solution for Ridge regression

$$W^* = (X^TX + \lambda I)^{-1}X^TY$$

Question 3

$VC(\mathcal{H})$ is the maximum cardinality of any set of instances that can be shattered by $\mathcal{H}$. We say that $\mathcal{H}$ shatters a set of points if and only if it can assign any possible labeling to those points.

1. We should show that the VC dimension $d_{\mathcal{H}}$ of any finite hypothesis space $\mathcal{H}$ is at most $log_2\mathcal{H}$.
   Proof:
   For any set of distinct points S of size n, there are 2^n distinct ways of labeling those points. This means that for $\mathcal{H}$ to shatter S it must contain at least 2^n distinct hypotheses. This tells us that if the VC dimension of $\mathcal{H}$ is n then we must have 2^n hypotheses, i.e. $2^n \leq |\mathcal{H}|$ or equivalently that $n = VC(\mathcal{H}) \leq log_2|\mathcal{H}|$.

2. Consider a domain with n binary features and binary class labels. Let $\mathcal{H}$ be the hypothesis space that contains all decision trees over those features that have depth no greater than d. (The depth of a decision tree is the depth of the deepest leaf node.)
   Proof:
   First note that any tree in $\mathcal{H}$ can be represented by a tree of exactly depth d in $\mathcal{H}$. So we will restrict our attention to trees of exactly depth d. All of these trees have 2^d leaf nodes. Also note that there are a total of 2^n examples in our instance space, which gives us an immediate upper bound on the VC-dimension of $\mathcal{H}$, i.e. $VC(\mathcal{H}) \leq 2^n$.
   To get a lower bound let S contain the set of all possible 2^n instances. Since we have that $d \geq n$ it is straightforward to create a tree of depth n with a leaf node for each example and furthermore we can label the leaf nodes in all possible ways. This shows that we can shatter the set S with $\mathcal{H}$, which implies that $VC(\mathcal{H}) \geq 2^n$. Combining the upper and lower bound tell us that $VC(\mathcal{H}) = 2^n$.
   Therefore, given some $d > 1$, we showed that a tight bound(theta bound) is possible such that $VC(\mathcal{H}) = d$ i.e. $d = d_H$.

Question 5

Let $\vec{x_1}, \vec{x_2}...\vec{x_n}$ be the feature vectors of n data points in the original feature space. Let ϕ be the feature tranformation function. Then, $\phi(\vec{x_1}), \phi(\vec{x_2})..., \phi(\vec{x_n})$ are the feature vectors in the transformed feature space.
Let K be the kernel function such that:

$$K(i,j) = \phi(x_i)^T\phi(x_j)$$

The center of mass, $\vec{\mu}$, in the feature space can be defined as the average of the

vectors in the transformed feature space.

$$\vec{\mu} = \frac{1}{n}\sum_{i=1}^{n}\phi(\vec{x_i})$$

Consider:

$$\begin{aligned}
||\mu||^2 &= \mu^T\mu \\
&= \mu^T\frac{1}{n}\sum_{i=1}^{n}\phi(\vec{x_i}) \\
&= \frac{1}{n}\sum_{j=1}^{n}\phi(\vec{x_j})^T\frac{1}{n}\sum_{i=1}^{n}\phi(\vec{x_i}) \\
&= \frac{1}{n^2}\sum_{i,j}\phi(\vec{x_j})^T)\phi(\vec{x_i}) \\
&= \frac{1}{n^2}\sum_{i,j}K(i,j)
\end{aligned}$$

Average of the squared Euclidean distances from μ to each $\phi(x)$

The squared euclidean distance of a single feature vector in the transformed space from the center of mass $\vec{\mu}$ can be expressed as follows:

$$\begin{aligned}
||\phi\vec{x_i}) - \vec{\mu}||^2 &= (\phi(\vec{x_i}) - \vec{\mu})^T(\phi(\vec{x_i}) - \vec{\mu}) \\
&= \phi(\vec{x_i})^T\phi(\vec{x_i}) - 2\phi(\vec{x_i})^T\vec{\mu} + ||\vec{\mu}||^2 \\
&= K(i,i) - \frac{2}{n}\phi(\vec{x_i})^T\sum_{j=1}^{n}\phi(\vec{x_j}) + ||\vec{\mu}||^2 \\
&= K(i,i) - \frac{2}{n}\sum_{j=1}^{n}\phi(\vec{x_i})^T\phi(\vec{x_j}) + ||\vec{\mu}||^2 \\
&= K(i,i) - \frac{2}{n}\sum_{j=1}^{n}K(i,j) + \frac{1}{n^2}\sum_{r,s}K(r,s)
\end{aligned}$$

The average of the euclidean distances of all the points from the center of mass can be written as:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}||\phi(\vec{x_i})-\vec{\mu}||^2 &= \frac{1}{n}\left(\sum_{i=1}^{n}\left(K(i,i)-\frac{2}{n}\sum_{j=1}^{n}K(i,j)+\frac{1}{n^2}\sum_{r,s}K(r,s)\right)\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}K(i,i)-\frac{2}{n}\sum_{i,j}K(i,j)+\frac{n}{n^2}\sum_{r,s}K(r,s)\right)\\
&= \frac{1}{n}\left(\sum_{i=1}^{n}K(i,i)-\frac{1}{n}\sum_{i,j}K(i,j)\right)
\end{aligned}$$

Thus, the average of euclidean distances from the center of mass $\vec{\mu}$ to each $\phi(x)$ can be expressed in terms of the kernel function K.