Machine Learning-I (CS/DS 864)

Midsem Exam

Prof.G.Srinivasaraghavan

Date Posted: Oct 4, 2016	**Submit By**: Oct 14, 2016, Midnight	**Max. Marks**: 100

Q-1: Consider the following scheme for 'composing' complex binary classification hypothesis classes from simple ones. Let $H_1, \ldots, H_k$ be k hypothesis classes with VC-Dimensions $\mathsf{d}_1, \ldots, \mathsf{d}_k$ where a hypothesis $h \in H_i$ for any i maps a vector $\boldsymbol{x} \in \mathcal{R}^d$ to ± 1. Let H_0 be a hypothesis class with VC-dimension d_0 that maps a vector $\boldsymbol{x} \in \{\pm 1\}^k$ to ± 1. Define a new hypothesis class H such that any hypothesis $h : \mathcal{R}^d \rightarrow \pm 1$ in H is of the form

$$h(\boldsymbol{x}) = h_0(h_1(\boldsymbol{x}), \ldots, h_k(\boldsymbol{x}))$$

where $\forall 0 \leq i \leq k, \;\; h_i \in H_i$. Let d_H be the VC-dimension of H and $D = \sum_{i=0}^{k} \mathsf{d}_i$. Show that

$$\mathcal{G}_H(n) \leq \prod_{i=0}^{k} \mathcal{G}_{H_i}(n) \;\; \text{and} \;\; \mathsf{d}_H \leq 2D \log_2(D)$$

whenever $D > e \log_2(D)$. **Max Marks: 20**

Q-2:

a. Show that all the eigenvalues of the Hat Matrix $X(X^TX)^{-1}X^T$ (the rows of X are the data vectors) are either 0 or 1. Assume X^TX is invertible.

b. Show that any symmetric matrix A can be written as UDU^T for a given dataset where D is a diagonal matrix with the eigenvalues of A as the diagonal elements and columns of U are the corresponding eigenvectors forming an orthogonal basis. Conclude that the trace of A is the sum of its eigenvalues.

c. How many of the eigenvalues of the hat matrix are 1?

Max Marks: 7

Q-3: Recall that a positive definite matrix A is one whose eigenvalues are all positive. An alternate characterization of positive definite matrices is that $\forall \boldsymbol{y} \neq \boldsymbol{0} : \boldsymbol{y}^T A \boldsymbol{y} > 0$. Show that X^TX is a positive definite matrix. Use this to arrive at an alternate derivation of the least squares regression solution $\boldsymbol{w} = \left(X^TX\right)^{-1} X^T\boldsymbol{y}$, where X is the matrix in which each data point occupies one row (with an extra 1 for homogeneous coordinates) and $\boldsymbol{y}$ is the vector of given values for the response variable. Assume X^TX is invertible. **Max Marks: 8**

Q-4: This problem is to show a bound on the generalization error for standard linear regression. As in the previous problem, let X be the matrix in which each data point occupies one row (with an extra 1 for homogeneous coordinates) and $\boldsymbol{y}$ the vector of given values for the response variable. Also recall that $X\left(X^TX\right)^{-1}X^T$ is the hat matrix $\hat{H}$ — again assume X^TX is invertible. Answer the following towards proving the bound.

a. Recall the probabilistic view where the response value y is of the form $g(\boldsymbol{x}) + \epsilon$ where $g(\boldsymbol{x})$ is the function we are trying to estimate and ϵ is the random noise with zero mean and variance σ^2. Suppose $g(\boldsymbol{x})$ is of the form $g(\boldsymbol{x}) = \boldsymbol{x}^T.\boldsymbol{w}^*$ — this is the linear regression

case. Show that the estimate $\hat{\boldsymbol{y}}$ for $\boldsymbol{y}$ using the linear regression solution is given by $\hat{\boldsymbol{y}} = X\boldsymbol{w}^* + \hat{H}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is the vector of noise terms for each of the training data points. Also show that the error $y - \hat{y}$ between the observed value y and the predicted value $\hat{y}$ at a test point $(\boldsymbol{x}, y)$ is given by $\left(\epsilon - \boldsymbol{x}^T \left(X^T X\right)^{-1} X^T \boldsymbol{\epsilon}\right)$ where ϵ is the noise term at the test point. **Max Marks: 5**

b. Show that for any square symmetric matrix A and a vector $\boldsymbol{x}$

$$\boldsymbol{x}^T A \boldsymbol{x} = tr\left(\boldsymbol{x}\boldsymbol{x}^T A\right) \qquad (tr \text{ is the trace of a matrix})$$

and that for a random matrix M, $E[tr(M)] = tr(E[M])$. **Max Marks: 2**

c. Argue that assuming the size of the dataset is n, if n is large enough, for a random test data point $(\boldsymbol{x}, y)$ from the domain D, with a high probability

$$n.E_D\left[\boldsymbol{x}\boldsymbol{x}^T\right].\left(X^T X\right)^{-1} = \left(1 + o\left(\sqrt{n}\right)\right) I$$

Max Marks: 5

d. Show that the expected risk of linear regression with the square error loss function, across all possible datasets of size n is

$$\sigma^2\left(1 + \frac{d+1}{n} + o\left(\frac{d+1}{\sqrt{n}}\right)\right)$$

thus showing that the expected error converges to just the variance of the noise as n grows large. Recall the bias-variance tradeoff discussion in the class. Expectation across all possible datasets is taken as expectation across all possible occurrences of the noise. **Max Marks: 18**

Q-5: Here's a typical practical scenario, but presented in a fairly general setting — we have k different learning algorithms $\mathcal{A}_1, \ldots, \mathcal{A}_k$ with $\mathcal{H}_1, \ldots, \mathcal{H}_k$ as the corresponding hypothesis classes each of which consists of hypotheses operating in an input domain $\mathcal{D}$. Another way to think of this is that we have one learning algorithm and each $\mathcal{A}_i$ corresponds to one set of choices for the hyperparameters — clearly the hyperparameter choices would determine the class of models that would be considered, resulting in the corresponding $\mathcal{H}_i$. Note that this means the algorithm $\mathcal{A}_i$ will train a model from the class $\mathcal{H}_i$. Assume we have a dataset D_n of size n and that we have split it into a training subset $T \subset D_n$ of size αn and a validation subset $V \subset D_n$ of size $(1-\alpha)n$ for some $0 < \alpha < 1$. Each algorithm $\mathcal{A}_i$ trains a model from $\mathcal{H}_i$ using the training dataset and produces the model $\hat{h}_i$ that minimizes the empirical error within the training set for some appropriately chosen loss function. Assume that all the algorithms use the same loss function. Let c_i denote the number of equivalence classes of models in $\mathcal{H}_i$ — we consider two hypotheses h_1, h_2 to be equivalent iff $h_1(D_n) \equiv h_2(D_n)$. Note that if the hypothesis class is finite then c_i is upper bounded by $|\mathcal{H}_i|$. We then compute the empirical loss for each model $\hat{h}_i$ on the validation set and pick the one with the least validation loss. Let the final model be $\hat{h}$. We want to prove a generalization bound for $\hat{h}$. Suppose

$$h^* = \arg \min_{h \in \{\mathcal{H}_1 \cup \ldots \cup \mathcal{H}_k\}} E_{\boldsymbol{x} \in \mathcal{D}}\left[\mathcal{L}(\boldsymbol{x}, h)\right]$$

— h^* is the model that would minimize the true risk across the domain $\mathcal{D}$; the best one among all the models in $\mathcal{H}_1 \cup \ldots \cup \mathcal{H}_k$. Suppose $h^* \in \mathcal{H}_l$ for some $1 \leq l \leq k$. Show that with probability at least $(1 - \delta)$

$$R(\hat{h}) \leq R\left(h^*\right) + \sqrt{\frac{2}{\alpha n} \log \frac{8c_l}{\delta}} + \sqrt{\frac{2}{(1-\alpha)n} \log \frac{8k}{\delta}}$$

Intuitively what we are saying is that using the hold-out strategy that we have used we will arrive at a model whose true risk is different from the true risk of the best possible model by a bounded error margin and that the error margin vanishes to zero as $n \to \infty$. This also reveals a tradeoff between the training and validation set sizes.

To prove this you can use the following form of the Azuma-Hoeffding inequality — for any iid sample $X_1, \ldots, X_n$ from a domain with mean μ

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) < 2e^{-2\epsilon^2 n}$$

Hint: Bound (i) $R(\hat{h})$ using $R_V(\hat{h})$ and in turn $R_V(\hat{h}_l)$, (ii) $R_V(\hat{h}_l)$ in terms of $R(\hat{h}_l)$, (iii) $R(\hat{h}_l)$ in terms of $R_T(\hat{h}_l)$ and in turn $R_T(h^*)$, and finally (iv) $R_T(h^*)$ using $R(h^*)$. R_T and R_V denote empirical errors within the training and validation data respectively. Put all these together to get the final bound. Follow a similar sequence of arguments we went through when we proved the VC theorem. **Max Marks: 25**

Note: 10 Marks grace for presentation, clarity, precision. 100 marks disgrace :-(for any attempt to copy, particularly of the blind, mindless kind!!