10-701/15-781, Machine Learning: Homework 4

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on March 24, 2010

1 Learning Theory [25pt, Ni Lao]

Sample complexity and model complexity are two important concepts in machine learning. We will explore them in this question, and you will also practice you skill in estimating VC dimensions.

1.1 Gaussian Bayes Model [15 pt]

Let's consider several Gaussian Bayes classification Models for the classification problem $P(Y|X)$. All models here assume there are two classes ($Y \in \{0, 1\}$), that X is a vector of real-valued features, and $P(X|Y = 1)$ and $P(X|Y = 0)$ are modeled by two different Gaussian distributions. Please complete the following table by filling in the number of parameters and VC dimension of different Gaussian Bayes learning models under different settings. Each column describes a different Bayesian classification model in terms of (1) the number of features in X, (2) whether the two Gaussians for $P(X|Y = 1)$ and $P(X|Y = 0)$ share the same co-variance matrix, and (3) whether the model makes the naive Bayes assumption of conditional independence among features. (note the first two columns have '-' for Naive Bayes because these columns assume X has just one feature).

ID	**a**	**b**	**c**	**d**	**e**	**f**
No. features	1	1	2	2	2	2
Shared Covariance Matrix?	Y	N	Y	Y	N	N
Naive Bayes?	-	-	Y	N	Y	N
No. parameters						
VC dimension						

Policy: number of parameters is 0.5pt each. VC dimension is 2pt each, and you get $\max(0, 2 - d_{best} + d_{your})$pt, where d_{best} is the best bound I know, and d_{your} is your answer. You need to convince me in order to get credit for the VC dimension, but you need not give a formal proof.

Hint: think about what kind of decision boundary we get in each of the models.

1.2 VC dimension and Effective Number of Hypothesis [3 pt]

For models where the set of instances X involve continuous variables, the hypothesis space H defined over X may contain an infinite number of hypothesis. In order to quantify the richness of our hypothesis space, we now define the Effective Number of Hypotheses of a model with respect to a set of unlabeled data points $D = (x_1, \cdots, x_m)$ as $N_{eff}(H, D)$, which is the number of different ways samples in D can be divided into positive and negative ones by hypotheses in H.

What is the relationship between $N_{eff}(H, D)$ and m, the number of samples in D?

Hint: discuss separately for $m \leq VC(H)$ and $m > VC(H)$

1.3 [2 pt]

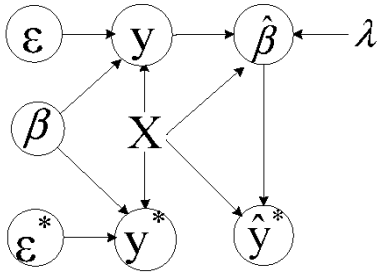
Assume no noise in the labels. At least how many labeled samples do we need in order to identify one out of the $N_{eff}(H, \mathcal{D})$ hypothesis? **Hint:** think about entropy (or information)

1.4 Linear Regression Model [3 pt]

Now let's explore the relations between sample complexity, model complexity, and number of parameters on a model we all know so well about.

In homework 3 we derived the expected error of linear (and ridge) regression. Basically, we are given training data of the form, $\mathcal{D} = (\mathbf{X}, \mathrm{y}) = \{(\mathrm{x}_i, y_i)\}, i = 1, 2, ..., n$, where $\mathrm{x}_i \in \mathcal{R}^{1\times p}$, i.e. $\mathrm{x}_i = (x_{i,1}, \cdots, x_{i,p})^{\mathrm{T}}$, $y_i \in \mathcal{R}$, $\mathbf{X} \in \mathcal{R}^{n\times p}$, where n is number of samples, p is number of features. Each row i of $\mathbf{X}$ is $\mathrm{x_i}^{\mathrm{T}}$, and $\mathrm{y} = (y_1, \cdots, y_n)^{\mathrm{T}}$. We assumed that $p < n$, and $\mathbf{X}^T\mathbf{X}$ is invertible. Also assume that our data is generated from a true model of the form: $y_i = \mathrm{x_i}^{\mathrm{T}}\beta + \epsilon_i$ (or in matrix form $\mathrm{y} = \mathbf{X}\beta + \epsilon$), where $\epsilon_1, ..., \epsilon_n$ are IID and sampled from a Gaussian with 0 mean and constant standard deviation, that is $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (or $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$).

We also assumed that the true parameter β itself is a random variable $\sim \mathcal{N}(0, \alpha^2 I)$. Apart from our training data $\mathcal{D} = (\mathbf{X}, \mathrm{y})$, we generate a set of testing data $\mathcal{D}^* = (\mathbf{X}, \mathrm{y}^*)$. It has exactly the same x values $\mathbf{X}$ as training data, but the y values are regenerated independently. Again we can decompose them as $\mathrm{y}^* = \mathbf{X}\beta + \epsilon^*$. Relations among these quantities can be summarized by the figure below.



Finally, we showed that the risk of ridge regression can be expressed in terms of the regularization parameter λ as

$$R(\lambda) = E[e(\lambda)^T e(\lambda)] \tag{1}$$

$$= \sum_{i=1..p} \left[\left(\frac{\lambda d_i}{d_i^2 + \lambda} \right)^2 \alpha^2 + \left(\frac{d_i^2}{d_i^2 + \lambda} \right)^2 \sigma^2 \right] + \sum_{i=1..n} \sigma^2, \tag{2}$$

Here we decompose X as $X = UDV^T$ by using SVD, where D is a $p \times p$ diagonal matrix, V is a $p \times p$ unitary matrix, U is a $n \times p$ matrix, which is the first p columns of a unitary matrix. We define $d_i = D_{i,i}$.

Please express the average risk $R(\lambda)/n$ for linear regression (not ridge regression) as a function of p and n. It has two terms, one corresponds to the irreducible error, one corresponds to the variance of the model.

Hint: the above solution to HW3 Q1 is slightly revised (see the online documents). Basically U should be $n \times p$ instead of $n \times n$

1.5 [2 pt]

How many training examples n suffice to assure that the error term which corresponds to the model variance will be no larger than the irreducible error? What is its relationship to the VC dimension of linear regression model?