COS 424: Interacting with Data

Lecturer: David Blei
Scribe: Christopher DeCoro

Lecture 16 (4/3/2008)

We are reviewing the discussion of bias-variance tradeoff that was introduced last week (in particular for the case of linear regression models, though the concepts apply generally). Consider the following model, in which the response variable y_n is modeled as a function of the variable x_n according to the rule:

$$y_n \sim N(\beta^T x_n, \sigma^2). \tag{1}$$

Note that $N(\mu, \sigma^2)$ is the normal distribution with mean μ and standard deviation σ, the input variable x_n is represented in homogeneous coordinates (padded with an additional dimension whose value is fixed to 1), and β is a vector of the same dimensionality as x_n.

Given a specific dataset $x_1...x_n$, the maximum likelihood estimate for the parameter $\hat{\beta}$, which is the choice of β in (1) that minimizes training error, can be computed as:

$$\hat{\beta} = \arg\min_{\beta} \log p(y_{1..n}|x_{1..n}, \beta) \tag{2}$$

We can thus conceptualize $\hat{\beta}$ as a random variable. Suppose for the purpose of our discussion that we know the "true"" β, and that we will pick the points x in advance. We can generate a set of response y by finding the points on the line $\beta^T x$, and use this point as the mean of a normal distribution from which to draw a random sample. Given these responses y, we can "discard" (for demonstration purposes) our knowledge of β, and compute the maximum-likelihood estimate $\hat{\beta}$ from the data. Because of the finite sample size, $\hat{\beta}$ will not be equal to the underlying value β. Instead, $\hat{\beta}$ is a random variable that is governed by a random process. We could retrieve an unlimited number of instantiations of the random variable $\hat{\beta}$ by resampling the dataset.

Suppose now that we have a new data point drawn from the actual distribution, $(x_0, y_0 = \beta^T x_0)$ (equivalently, the conditional expectation of y_0 given x_0, $E\left[y_0|x_0\right]$, is βx_0). For a given estimated parameter $\hat{\beta}$ computed from our training dataset, the estimated response $\hat{y_0}$ of x_0 is $\hat{\beta}^T x_0$. We can now contemplate the mean-squared error (MSE) of the prediction:

$$E\left[(\hat{y}_0 - y_0)^2\right] = (\hat{\beta}x_0 - \beta x_0)^2 \tag{3}$$

The previous assumes that we have fixed $\hat{\beta}$, and measures the prediction of $\hat{y}$ relative to the true y; when considering $\hat{\beta}$ as an additional random variable, we can also express the error in the estimator's prediction of β, for all possible choices of the data.

$$\begin{aligned} MSE(\hat{\beta}x_0) &= E_{\mathcal{D}}\left[\left(\hat{\beta}(\mathcal{D})x_0 - \beta x_0\right)^2\right] & (4)\\ &= E[(\hat{\beta}x_0)^2] - 2E[\hat{\beta}x_0]\beta x_0 + (\beta x_0)^2 + E[\hat{\beta}x_0]^2 - E[\hat{\beta}x_0]^2 & (5)\\ &= \left(E[(\hat{\beta}x_0)^2] - E[\beta x_0]^2\right) + (E[\beta x_0] - \beta x_0)^2 & (6) \end{aligned}$$

Note that in (4) we explicitly show that $\hat{\beta}$ is a function of the data $\mathcal{D}$ over which the expectation is computed, a notation we drop for simplicity in subsequent expressions. We expand (4) using linearity of expectation, and add zero to produce (5). Combining terms, we produce (6), the left term of which is the *variance* of the estimator, and the right term is the *squared bias.*

Considering only the squared bias term, the expression $E[\beta x_0]$ is the average value of the mean when estimated over all possible datasets $\mathcal{D}$ drawn from the underlying distribution; the expression βx_0 is the true mean of that distribution. Originally, statistics was solely concerned with *unbiased* estimators, in which the bias term is zero. In particular, the MLE is the unbiased estimator with minimum variance (this statement is known as the Gauss-Markov theorem). Ultimately, however, we care about prediction error; as (6) indicates that error is a function of both bias and variance, we might consider using a biased estimator that significantly reduces variance relative to the MLE.

One common class of such methods are known under the general term of *regularization*, in which constraints are placed on the potential values of $\hat{\beta}$. This encourages smaller and simpler models, with less possible variance. In particular, the regularization method known as *ridge regression* (a type of *shrinkage*) minimizes mean-squared prediction error subject to constraints on the norm of β

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{n} \frac{1}{2}(y_n - \beta x_n)^2 + \lambda \sum_{i} \beta_i^2 \tag{7}$$

¿From examination, the meta-parameter λ acts as a weight that controls the contribution of the L_2 norm of the vector β (correspondingly, its distance to origin in parameter space) to the total error; with $\lambda = 0$, the minimization of prediction error is unconstrained, as λ increases the minimization is forced to choose values of β with smaller norm (or closer to the origin, and thereby constraining the range of possible values). The minimization problem can be shown to remain convex, and therefore has a unique global minimum, for fixed λ. The value of λ is found from the data by the method of *cross-validation*, according to the following steps:

1. Partition dataset into k subsets, or *folds*
2. Decide on candidate values of λ
3. For each combination of fold and candidate λ, estimate β on out- of-fold samples.
4. Compute corresponding prediction error for in-fold samples
5. Choose λ that minimizes error

In conclusion, the concept of bias-variance tradeoff is fundamental to the Bayesian approach to statistics: by assuming a prior distribution on the data and parameters, we depart from the MLE, resulting in an estimator that is inherently biased, although in practice frequently performs significantly better than unbiased estimators.