COS 424: Interacting with Data

Lecturer: David Blei
Scribe: Jeehyung Lee

Lecture 4/3

---

1 Squared bias and variance of estimates

Given data $(x_{1:N}, y_{1:N})$,

MLE $\hat{B} = arg_B(max(log(y_{1:N}|x_{1:N}, B)))$

Suppose we know a true value B of some data. Suppose we sample random data using true value B based on Gaussian distribution. Then, the estimate $\hat{B}$ based on this data is not necessarily B.

Now suppose we observe a new input/output pair (x_o, y_o). The squared error of the estimate of y_o is,

$(\hat{B}x_o - Bx_o)^2$

This value indicates how close the prediction of the estimate $\hat{B}$ to that of the true B.

Considering $\hat{B}$ as a random variable, we can estimate Mean Squared Error of the estimates of y_o (Let $E_D[\hat{B}]$ denote $E[\hat{B}(D)]$, where D is a distribution of data from which $\hat{B}$ is estimated).

$MSE(\hat{B}x_o) = E_D[(\hat{B}x_o - Bx_o)^2]$

We can expand this equation to,

$MSE(\hat{B}x_o) = E_D[(\hat{B}x_o)^2] - 2E_D[\hat{B}x_o]Bx_o + (Bx_o)^2$
(Remember that B and x_o are fixed values here)

We add zero term $(E_D[\hat{B}x_o]^2 - E_D[\hat{B}x_o]^2)$ to this equation,

$MSE(\hat{B}x_o) = E_D[(\hat{B}x_o)^2] - 2E_D[\hat{B}x_o]Bx_o + (Bx_o)^2 + E_D[\hat{B}x_o]^2 - E_D[\hat{B}x_o]^2$

Now the equation is equivalent to

$(E_D[\hat{B}x_o] - Bx_o)^2 + (E[(\hat{B}x_o)^2] - E[\hat{B}x_o]^2)$

The first term is a "squared bias $(Bias^2(\hat{B}))$" and the second term is a "variance of estimates $(Var(\hat{B}))$".

According to Gauss-Markov Theorem, MLE is the unbiased estimator with the smallest variance. In other words, if $\hat{B}$ is a MLE, the squared bias will be 0 and the variance will be the smallest.

The prediction error, which is defined by the following equation,

$E_D[E_{y_o}[(\hat{B}x_o - y_o)^2]]$

is equal to,

$\delta^2 + Var(\hat{B}) + Bias^2(\hat{B})$

where δ^2 is the variance of data ($y_o \sim N(Bx_o, \delta^2)$).

2 Regularization

The basic idea of regularization is to trade $Var(\hat{B})$ and $Bias^2(\hat{B})$ by placing constraints on $\hat{B}$. This has 3-fold advantages,

- Encourages smaller and simpler models
- Makes the model robust to overfitting
- Makes the model more interpretable

One way to do this is Ridge Regression - to optimize RSS subject to constraint s on squared sum of coefficients. As s becomes bigger, we have a better chance of reducing error, but suffer a bigger variance.

$\hat{B}$ of Ridge Regression can be calculated by solving the following equation,

$\hat{B}^{ridge} = arg_B(min(\sum_{i=1}^{N} \frac{1}{2}(y_n - Bx_n)^2 + \lambda \sum_{i=1}^{p} B^2))$
(The term λ determines the size of the "Ball" constraining B)

This is a convex problem and can be solved efficiently.

As for the choice of λ, we choose λ from cross-validations to minimize test error - For candidate values of λ (i.e, grid between $0 \sim 1$) and for each fold, calculate $\hat{B}^{ridge}$ and get average error of within-fold samples. We choose λ that minimizes the average error.

3 Bayesian Statistics

Parameter $\theta \sim G_o(\alpha)$

$y_n \sim F(\theta)$

Posterior $p(\theta|y_{1:N}, \alpha)$

where $G_o(\alpha)$ is a prior distribution and α is called "hyperparameter".

To calculate MLE, Bayesians choose θ to maximize likelihood of $y_{1:N}$. Bayesian estimates give up on bias to reduce variance.