COS 513 - October 18 Scribe Notes

Manish Nag

October 26, 2010

1 Linear regression

Let $\mathscr{D} = \{(X_n, Y_n)\}_{n=1}^N$. Find β that can predict Y_{new} from X_{new}

Fit coefficients to minimize the sum of squared distances from a line to all points. The objective function of β, a scalar, is:

$$RSS(\beta) = \frac{1}{2}\sum_{n=1}^{N}(Y_n - \beta X_n)^2$$

RSS: Residual sum of squares
To optimize on β, we take the derivative.

$$\frac{d\,RSS(\beta)}{d\,\beta} = -\sum_{n=1}^{N}(Y_n - \beta X_n)X_n$$

The optimal coefficient $\hat{\beta}$ can be shown as:

$$\hat{\beta} = \frac{\sum_{n=1}^{N} Y_n X_n}{\sum_{n=1}^{N} X_n^2}$$

$$\hat{Y}_{new} = \hat{\beta} X_{new}$$

Note: X_{new} is always assumed as given.
In general:

$$Y = \boldsymbol{\beta}_0 + \sum_{i=1}^{P}\boldsymbol{\beta}_i X_{n,i}$$

where $\boldsymbol{\beta}$ is assumed to be a vector of n coefficients and an intercept term. To simplify, set

$$\boldsymbol{\beta}_{p+1} = \boldsymbol{\beta}_0\,, X_{p+1} \triangleq 1$$

$$y = \boldsymbol{\beta}^T X$$

and

$$RSS(\boldsymbol{\beta}) = \frac{1}{2}\sum_{n=1}^{N}(Y_n - \boldsymbol{\beta}^T X_n)^2$$

The gradient

$$\nabla RSS(\boldsymbol{\beta}) = -\sum_{n=1}^{N}(Y_n - \boldsymbol{\beta}^T X_n)X_n$$

Here, we could use an optimizer, but for linear regression we can solve this exactly.

The design matrix is defined as

$$X = \begin{bmatrix} X_{11}, & X_{12}, & ... & X_{1p} \\ X_{21}, & X_{22}, & ... & X_{2p} \\ X_{n1}, & X_{n2}, & ... & X_{np} \end{bmatrix}$$

The response vector is $Y =< Y_1, ..., Y_n >$ where each component is a response from the observed data.

$$lmR = lm(y \sim x)$$

$$\nabla_{\boldsymbol{\beta}} RSS(\boldsymbol{\beta}) = -X^T(Y - X\boldsymbol{\beta})$$

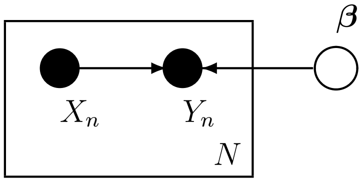
set to $\emptyset$ vector:

$$X^T Y - X^T X\hat{\boldsymbol{\beta}} = 0$$

$$\hat{\boldsymbol{\beta}} = \frac{X^T Y}{X^T X}$$

called normal equations. Note: $X^T X$ needs to be invertible.

2 Probabilistic interpretation



Conditional Model

$$Y_n \,|\, X_n \sim N(\boldsymbol{\beta}^T X_n, \sigma^2)$$

Constant σ^2 indicates that a Gaussian bump exists at each point along the prediction line, with a mean being the predicted mean and variance σ^2.

Fit $\boldsymbol{\beta}$ conditional maximum likelihood

$$\mathscr{L}(\boldsymbol{\beta}\,;\mathscr{D}) = log \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}}\; exp\left\{-\;\frac{1}{2\sigma^2}(Y_n - \boldsymbol{\beta}^T X_n)^2\right\}$$

$$= \sum_{n=1}^{N} -\frac{1}{2}\, log\, 2\pi\sigma^2 - \frac{1}{2\sigma^2}(Y_n - \boldsymbol{\beta}^T X_n)^2$$

Optimize w/r/t $\boldsymbol{\beta}$, same as minimizing the RSS.

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(Y_n - \boldsymbol{\beta}^T X_n)^2$$

Prediction: $\mathbb{E}\,[Y_{new}|X_{new}\,, \boldsymbol{\beta}] = \hat{\boldsymbol{\beta}}^T X_{new}$
Note:σ^2 doesn't play a role.

3 Bias-variance tradeoff

Consider a random data set drawn from a linear regression model. Hold the design matrix fixed and draw

$$Y_n|X_n\,, \boldsymbol{\beta} \sim N(\boldsymbol{\beta}^T X_n, \sigma^2)$$

Contemplate $\hat{\boldsymbol{\beta}}$ as a random variable governed by prediction by the distribution of the data. Consider a new input X. How does the prediction from $\hat{\boldsymbol{\beta}}$ compare to the "true" prediction $\boldsymbol{\beta}^T X$?

$$MSE(\hat{\boldsymbol{\beta}}^T X) = \mathbb{E}_{\mathscr{D}}\left[(\hat{\boldsymbol{\beta}}^T X - \boldsymbol{\beta}^T X)^2\right]$$

expanding:

$$= \mathbb{E}[(\hat{\boldsymbol{\beta}}^T X)^2] - 2\mathbb{E}[\hat{\boldsymbol{\beta}}^T X]\boldsymbol{\beta}^T X + (\boldsymbol{\beta}^T X)^2 + \mathbb{E}[\hat{\boldsymbol{\beta}}^T X]^2 - \mathbb{E}[\hat{\boldsymbol{\beta}}^T X]^2$$

$$= \mathbb{E}[(\hat{\boldsymbol{\beta}}^T X)^2] - \mathbb{E}[\hat{\boldsymbol{\beta}}^T X]^2 + (\mathbb{E}[\hat{\boldsymbol{\beta}}^T X] - \boldsymbol{\beta}^T X)^2$$

$(\mathbb{E}[\hat{\boldsymbol{\beta}}^T X] - \boldsymbol{\beta}^T X)^2$ is the squared bias. Variance reflects how sensitive the estimate is to randomness inherent in the data.

If $(\mathbb{E}[\hat{\boldsymbol{\beta}}^T X] - \boldsymbol{\beta}^T X) = 0$, then the estimator is unbiased. Classical statistics focused solely on unbiased estimators. At present, some bias is considered is the variance is lowered.

4 Gauss-Markov Theorem

The MLE/least squares estimate is the unbiased estimate with the least variance. In regression we trade off bias for variance through regularization (placing constraints on $\boldsymbol{\beta}$).

If the true MLE $\boldsymbol{\beta}$ lives outside the constraints, then the estimate must be biased. Regularization encourages smaller and simpler models that are easier to try to interpret. This approach is more robust to overfitting, which results in poor predictive power due to matching too closely to a training data set.