10601 Machine Learning

October 12, 2011

Mladen Kolar

Outline

- Bias – Variance tradeoff
- Linear regression
- Bayes networks

BIAS – VARIANCE TRADEOFF

Applet for least squares

http://mste.illinois.edu/users/exner/java.f/leastsquares/

Decomposition of error

Assume $Y = f(x) + \epsilon$

Generalization error

$$\mathrm{err}(x_0) = E[(Y - \hat{f}(X))^2 | X = x_0]$$

$$\mathrm{err}(x_0) = \underbrace{\sigma^2}_{\text{unavoidable error}} + \underbrace{(E_{\mathcal{D}}[\hat{f}(x_0) - f(x_0)])^2}_{\text{bias}} + \underbrace{Var_{\mathcal{D}}(\hat{f}(x_0))}_{\text{variance}}$$

Bias

Suppose that we have multiple datasets with n samples

On each data set we learn $\hat{f}(x)$

On average (over different datasets) we learn $E[\hat{f}(x)]$

Bias measures the difference between what you expect to learn and the truth

- decreases with complexity of the model

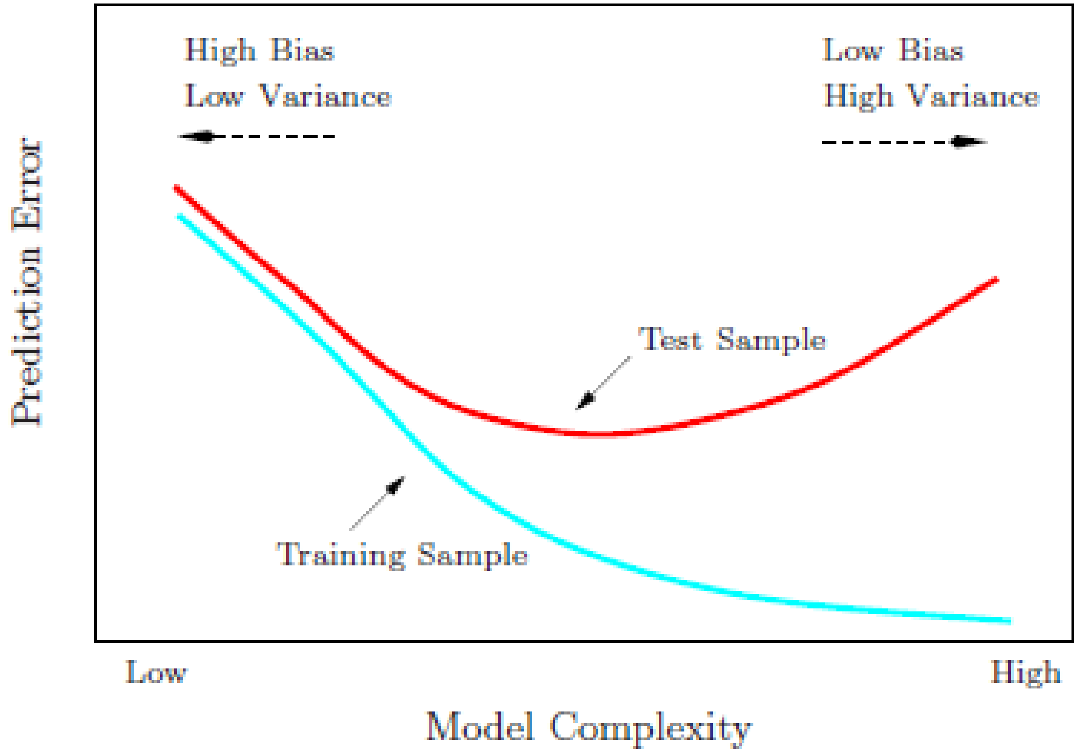
Variance

Measures the difference between what you expect to learn and what you learn on a particular dataset.

Measures how sensitive learner is to a specific dataset

Decreases as we have simpler models

Model complexity



Hastie, Tibshirani, Friedman

LINEAR REGRESSION

Linear regression model

$$y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

$$p(y|x) = N(f(x), \sigma^2)$$

$$f(x) = w_0 + \sum_i w_i \phi_i(x)$$

Maximum conditional likelihood estimation

$$\hat{w} = \arg\min_w \sum_l (y_l - \sum_i w_i \phi_i(x_l))^2$$

Matrix of transformed features

$$\Phi = \begin{pmatrix} \phi_1[X_{11}, \ldots, X_{1\,d}] & \ldots & \phi_m[X_{11}, \ldots, X_{1\,d}] \\ \ldots & \ldots & \ldots \\ \phi_1[X_{\mathrm{n}1}, \ldots, X_{\mathrm{nd}}] & \ldots & \phi_m[X_{\mathrm{n}1}, \ldots, X_{\mathrm{nd}}] \end{pmatrix}$$

$$\Phi = \begin{pmatrix} X_{11} & \ldots & X_{1\,d} \\ \ldots & \ldots & \ldots \\ X_{\mathrm{n}1} & \ldots & X_{\mathrm{nd}} \end{pmatrix} = X$$

Linear regression (matrix equation)

$$y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} f_w(< X_{11}, \dots, X_{1d} >) + \epsilon_1 \\ \dots \\ f_w(< X_{n1}, \dots, X_{nd} >) + \epsilon_n \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{d} w_j X_{1j} + \epsilon_1 \\ \dots \\ \sum_{j=1}^{d} w_j X_{nj} + \epsilon_n \end{pmatrix} = X w + \epsilon$$

$$\hat{w} = \arg\min_{w} (y - Xw)'(y - Xw)$$

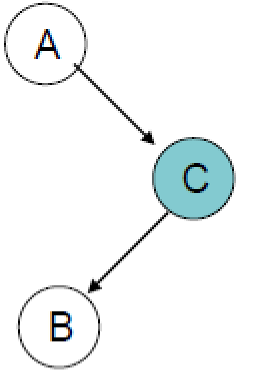
Final solution

$$\hat{w} = (X'X)^{-1}X'y$$

BAYES NETS

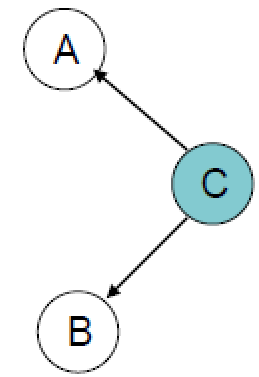
Head to Tail

$P(a,b|c) = P(a|c)P(b|c)$



Tail to Tail

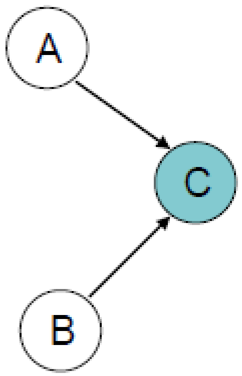
$P(a,b|c) = P(a|c)P(b|c)$



Where have we seen this?

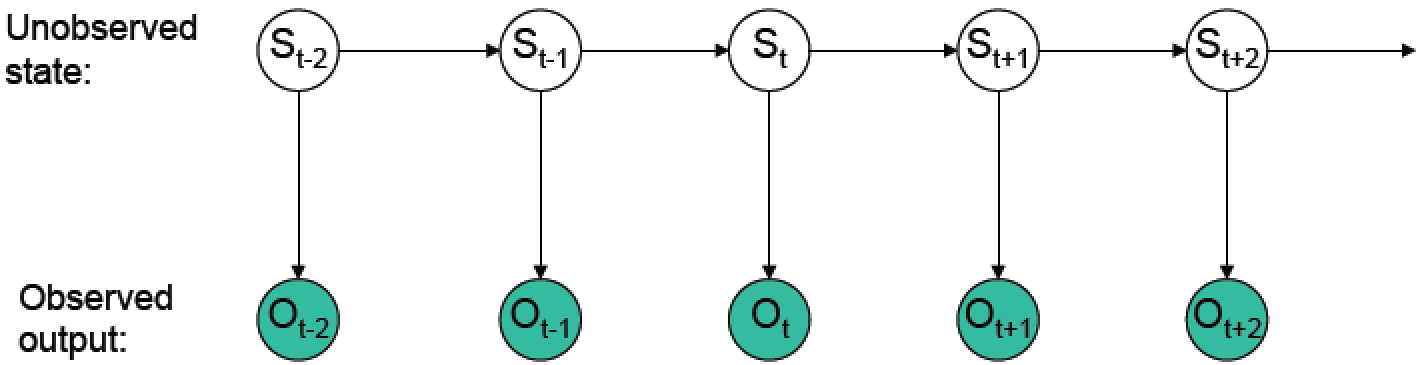
Head to head

$P(a,b) = P(a)P(b)$



What is the Bayes Network for $X_1, \ldots, X_4$ with no assumed conditional independencies?

Bayes Network for a Hidden Markov Model



$$P(S_{t-2}, O_{t-2}, S_{t-1}, \ldots, O_{t+2}) =$$

Implies the future is conditionally independent of the past, given the present

Bias – Variance tradeoff in Bayes nets

Give an example of very biased Bayes network?

Network with no edges

Naïve Bayes

Give an example of a network that has high variance?

Network of a distribution with no conditional independence assumptions

Questions?