

CS485/685

Lecture 5: Jan 17, 2012

Linear Regression by Maximum
Likelihood, Maximum A Posteriori and
Bayesian Learning
[B]: Sections 3.1 – 3.3

CS485/685 (c) 2012 P. Poupart

1

Noisy Linear Regression

- Assume y is obtained from $\mathbf{x}$ by a deterministic function f that has been perturbed (i.e., noisy measurement)

$$y = f(\bar{\mathbf{x}}) + \epsilon$$

$\downarrow$
 $\mathbf{w}^T \bar{\mathbf{x}}$

$\searrow$
 $N(0, \sigma^2)$

- Gaussian noise:

$$\begin{aligned} \Pr(\mathbf{y} | \bar{\mathbf{X}}, \mathbf{w}, \sigma) &= N(\mathbf{y} | \mathbf{w}^T \bar{\mathbf{X}}, \sigma^2) \\ &= \prod_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2}} \end{aligned}$$

CS485/685 (c) 2012 P. Poupart

2

Maximum Likelihood

- Possible objective: find best $\mathbf{w}^*$ by maximizing the likelihood of the data

$$\begin{aligned}\mathbf{w}^* &= \operatorname{argmax}_{\mathbf{w}} \Pr(\mathbf{y}|\bar{\mathbf{X}}, \mathbf{w}, \sigma) \\ &= \operatorname{argmax}_{\mathbf{w}} \prod_n e^{-\frac{(y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2}} \\ &= \operatorname{argmax}_{\mathbf{w}} \sum_n -\frac{(y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2} \\ &= \operatorname{argmin}_{\mathbf{w}} \sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2\end{aligned}$$

- We arrive at the original least square problem!

CS485/685 (c) 2012 P. Poupart

3

Maximum A Posteriori

- Alternative Objective: find $\mathbf{w}^*$ with highest posterior probability
- Consider Gaussian prior: $\Pr(\mathbf{w}) = N(\mathbf{0}, \Sigma)$
- Posterior:

$$\begin{aligned}\Pr(\mathbf{w}|\mathbf{X}, \mathbf{y}) &\propto \Pr(\mathbf{w}) \Pr(\mathbf{y}|\mathbf{X}, \mathbf{w}) \\ &= k e^{-\frac{\mathbf{w}^T \Sigma^{-1} \mathbf{w}}{2}} e^{-\frac{\sum_n (y_n - \mathbf{w}^T \mathbf{x}_n)^2}{2\sigma^2}}\end{aligned}$$

CS485/685 (c) 2012 P. Poupart

4

Maximum A Posterior

- Optimization:

$$\begin{aligned}\mathbf{w}^* &= \operatorname{argmax}_{\mathbf{w}} \Pr(\mathbf{w}|\bar{\mathbf{X}}, \mathbf{y}) \\ &= \operatorname{argmax}_{\mathbf{w}} - \sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2 - \mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w} \\ &= \operatorname{argmin}_{\mathbf{w}} \sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2 + \mathbf{w}^T \boldsymbol{\Sigma}^{-1} \mathbf{w}\end{aligned}$$

- Let $\boldsymbol{\Sigma}^{-1} = \lambda \mathbf{I}$ then

$$= \operatorname{argmin}_{\mathbf{w}} \sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2 + \lambda \|\mathbf{w}\|_2^2$$

- We arrive at the original **regularized** least square problem!

CS485/685 (c) 2012 P. Poupart

5

Expected Squared Loss

- Even though we use a statistical framework, it is interesting to evaluate the expected squared loss

$$\begin{aligned}E[L] &= \int_{\mathbf{x}, \mathbf{y}} \Pr(\mathbf{x}, \mathbf{y}) (y - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x} d\mathbf{y} \\ &= \int_{\mathbf{x}, \mathbf{y}} \Pr(\mathbf{x}, \mathbf{y}) (y - f(\mathbf{x}) + f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x} d\mathbf{y} \\ &= \int_{\mathbf{x}, \mathbf{y}} \Pr(\mathbf{x}, \mathbf{y}) \left[(y - f(\mathbf{x}))^2 + \underbrace{2(y - f(\mathbf{x}))(f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})}_{\text{Expectation is 0}} + (f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 \right] d\mathbf{x} d\mathbf{y}\end{aligned}$$

Expectation is 0

$$E[L] = \underbrace{\int_{\mathbf{x}, \mathbf{y}} \Pr(\mathbf{x}, \mathbf{y}) (y - f(\mathbf{x}))^2 d\mathbf{x} d\mathbf{y}}_{\text{noise (constant)}} + \underbrace{\int_{\mathbf{x}} \Pr(\mathbf{x}) (f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x}}_{\text{error (depends on } \mathbf{w} \text{)}}$$

CS485/685 (c) 2012 P. Poupart

6

Expected Squared Loss

- Let's focus on the error part, which depends on $\mathbf{w}$

$$E_{\mathbf{x}}[(f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2] = \int_{\mathbf{x}} \text{Pr}(\mathbf{x}) (f(\mathbf{x}) - \mathbf{w}^T \bar{\mathbf{x}})^2 d\mathbf{x}$$

- But the choice of $\mathbf{w}$ depends on the dataset S
- Instead consider expectation with respect to S

$$E_S[(f(\mathbf{x}) - \mathbf{w}_S^T \bar{\mathbf{x}})^2]$$

where $\mathbf{w}_S$ is the weight vector obtained based on S

CS485/685 (c) 2012 P. Poupart

7

Bias-Variance Decomposition

- Decompose squared loss

$$\begin{aligned} E_S[(f(\mathbf{x}) - \mathbf{w}_S^T \bar{\mathbf{x}})^2] &= E_S[f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] + E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}}]^2 \\ &= E_S \left[(f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}])^2 \right. \\ &\quad \left. + 2(f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}]) (E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}}) \right. \\ &\quad \left. + (E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}})^2 \right] \quad \text{Expectation is 0} \\ &= \underbrace{(f(\mathbf{x}) - E_S[\mathbf{w}_S^T \bar{\mathbf{x}}])^2}_{\text{bias}^2} + \underbrace{E_S[(E_S[\mathbf{w}_S^T \bar{\mathbf{x}}] - \mathbf{w}_S^T \bar{\mathbf{x}})^2]}_{\text{variance}} \end{aligned}$$

CS485/685 (c) 2012 P. Poupart

8

Bias-Variance Decomposition

- Hence:

$$E[\text{loss}] = (\text{bias})^2 + \text{variance} + \text{noise}$$

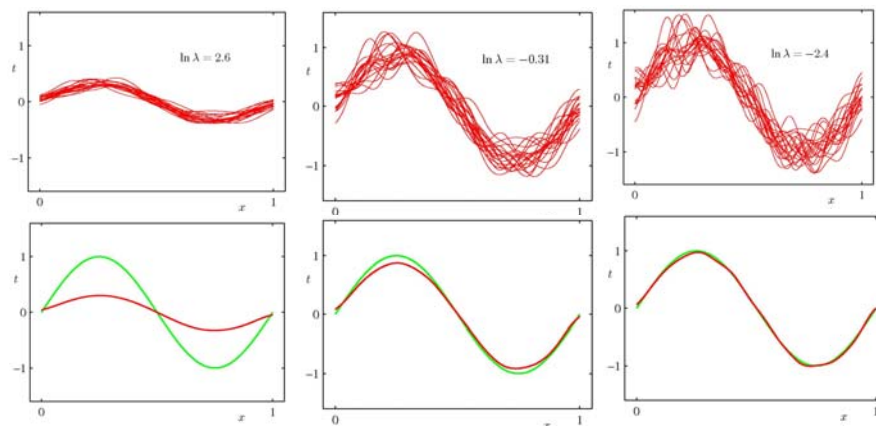
- Picture:

CS485/685 (c) 2012 P. Poupart

9

Bias-Variance Decomposition

- Example



CS485/685 (c) 2012 P. Poupart

10

Bayesian Linear Regression

- We don't know if $\mathbf{w}^*$ is the true underlying $\mathbf{w}$
- Instead of making predictions according to $\mathbf{w}^*$, compute the weighted average prediction according to $\Pr(\mathbf{w}|\bar{\mathbf{X}}, \mathbf{y})$

$$\Pr(\mathbf{w}|\bar{\mathbf{X}}, \mathbf{y}) = k e^{-\frac{\mathbf{w}^T \Sigma^{-1} \mathbf{w}}{2}} e^{-\frac{\sum_n (y_n - \mathbf{w}^T \bar{\mathbf{x}}_n)^2}{2\sigma^2}}$$

$$= k e^{-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A}(\mathbf{w} - \bar{\mathbf{w}})} = N(\bar{\mathbf{w}}, \mathbf{A}^{-1})$$

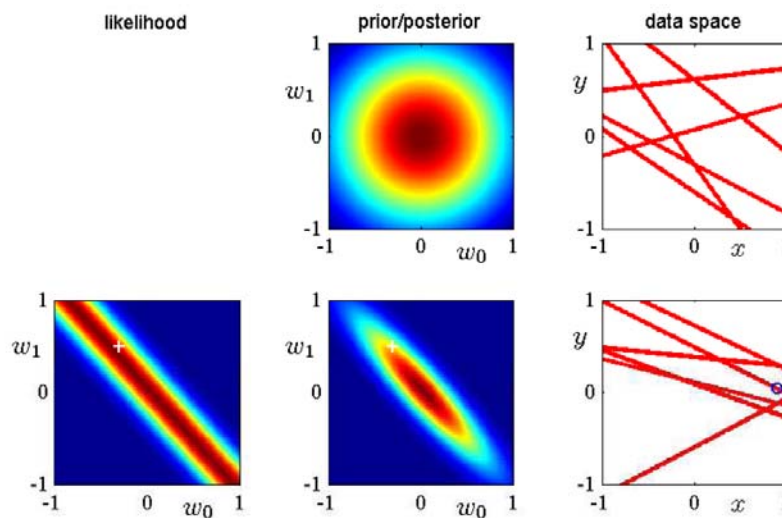
$$\text{where } \bar{\mathbf{w}} = \sigma^{-2} \mathbf{A}^{-1} \bar{\mathbf{X}}^T \mathbf{y}$$

$$\mathbf{A} = \sigma^{-2} \bar{\mathbf{X}}^T \bar{\mathbf{X}} + \Sigma^{-1}$$

CS485/685 (c) 2012 P. Poupart

11

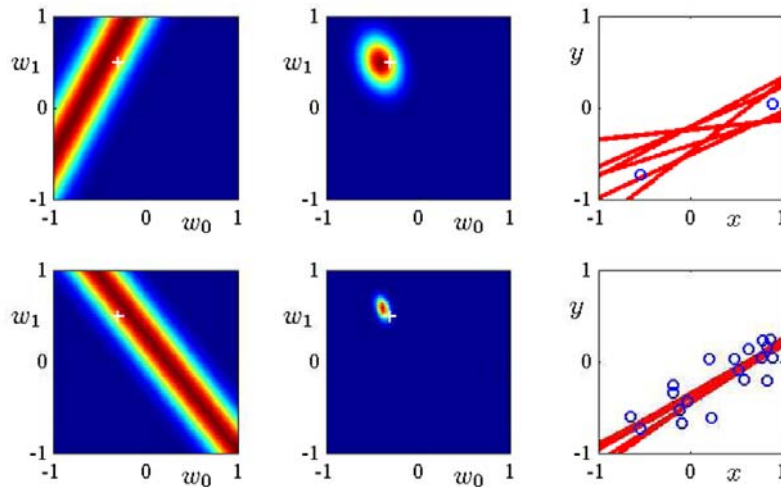
Bayesian Learning



CS485/685 (c) 2012 P. Poupart

12

Bayesian Learning



CS485/685 (c) 2012 P. Poupart

13

Bayesian Prediction

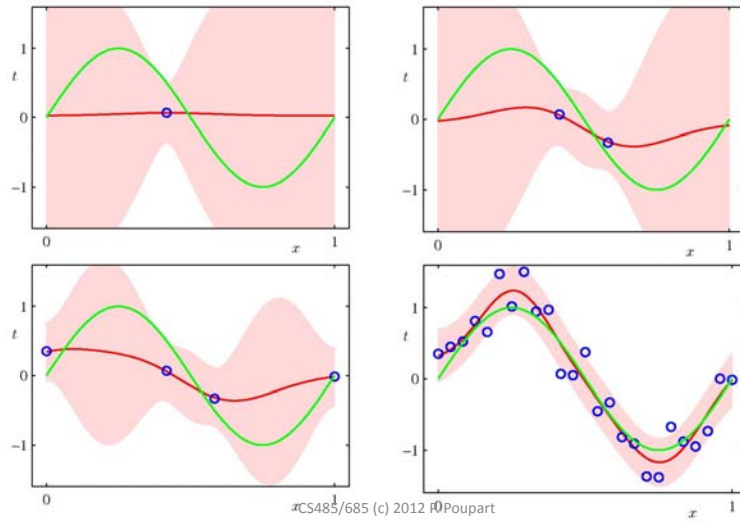
- Let x_* be the input for which we want a prediction and y_* be the corresponding prediction

$$\begin{aligned}
 \Pr(y_* | \bar{x}_*, \bar{X}, y) &= \int_{\mathbf{w}} \Pr(y_* | \bar{x}_*, \mathbf{w}) \Pr(\mathbf{w} | \bar{X}, y) d\mathbf{w} \\
 &= k \int_{\mathbf{w}} e^{-\frac{(y_* - \bar{x}_*^T \mathbf{w})^2}{2\sigma^2}} e^{-\frac{1}{2}(\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A}(\mathbf{w} - \bar{\mathbf{w}})} d\mathbf{w} \\
 &= N(\bar{x}_*^T \mathbf{A}^{-1} \bar{X}^T y, \bar{x}_*^T \mathbf{A}^{-1} \bar{x}_*)
 \end{aligned}$$

CS485/685 (c) 2012 P. Poupart

14

Bayesian Prediction



CS485/685 (c) 2012 J. Poupart

15