Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml2e

CHAPTER 14:

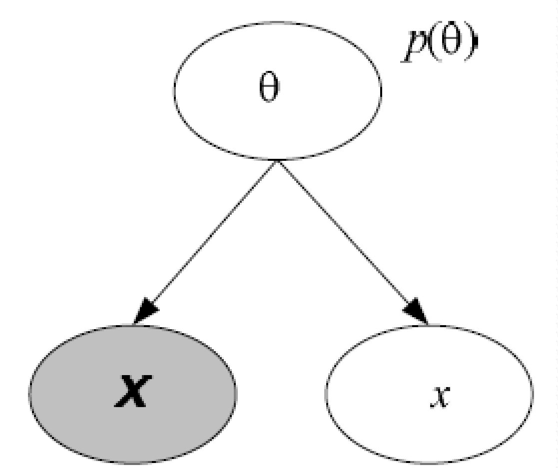
Bayesian Estimation

Rationale

- Bayes' Rule:

$$p(\theta \mid \mathrm{X}) = \frac{p(\theta)p(\mathrm{X} \mid \theta)}{p(\mathrm{X})}$$

- Generative model:



Estimating the Parameters of a Distribution: Discrete case

- x^t_i=1 if in instance t is in state i, probability of state i is q_i
- Dirichlet prior, α_i are hyperparameters

$$Dirichlet(\mathbf{q} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{i=1}^{K} q_i^{\alpha_i - 1}$$

- Sample likelihood

$$p(X \mid \mathbf{q}) = \prod_{t=1}^{N} \prod_{i=1}^{K} q_i^{x_i^t}$$

- Posterior

$$p(\mathbf{q} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + N_1)\cdots\Gamma(\alpha_K + N_K)} \prod_{i=1}^{K} q_i^{\alpha_i + N_i - 1}$$

$$= Dirichlet(\mathbf{q} \mid \boldsymbol{\alpha} + \mathbf{n})$$

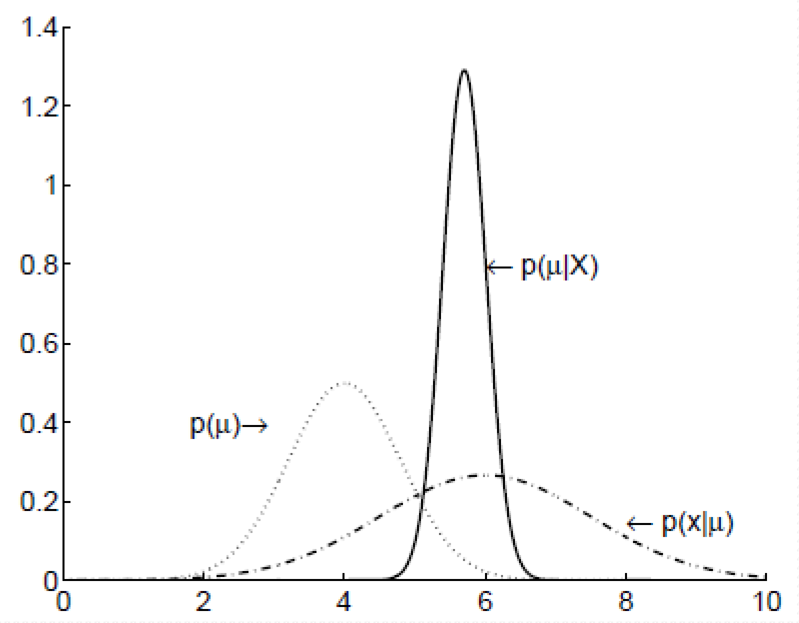
- Dirichlet is a conjugate prior
- With K=2, Dirichlet reduced to Beta

Estimating the Parameters of a Distribution: Continuous case

- $p(x^t)$~$N(\mu,\sigma^2)$
- Gaussian prior for μ, $p(\mu)$~ $N(\mu_0, \sigma_0^2)$
- Posterior is also Gaussian $p(\mu|X)$~ $N(\mu_N, \sigma_N^2)$ where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m$$

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$



Estimating the Parameters of a Function: Regression

- $r = \boldsymbol{w}^T\boldsymbol{x} + \varepsilon$ where $p(\varepsilon) \sim N(0, 1/\beta)$, and $p(r^t | x^t, \boldsymbol{w}, \beta) \sim N(\boldsymbol{w}^T\boldsymbol{x}^t, 1/\beta)$
- Log likelihood

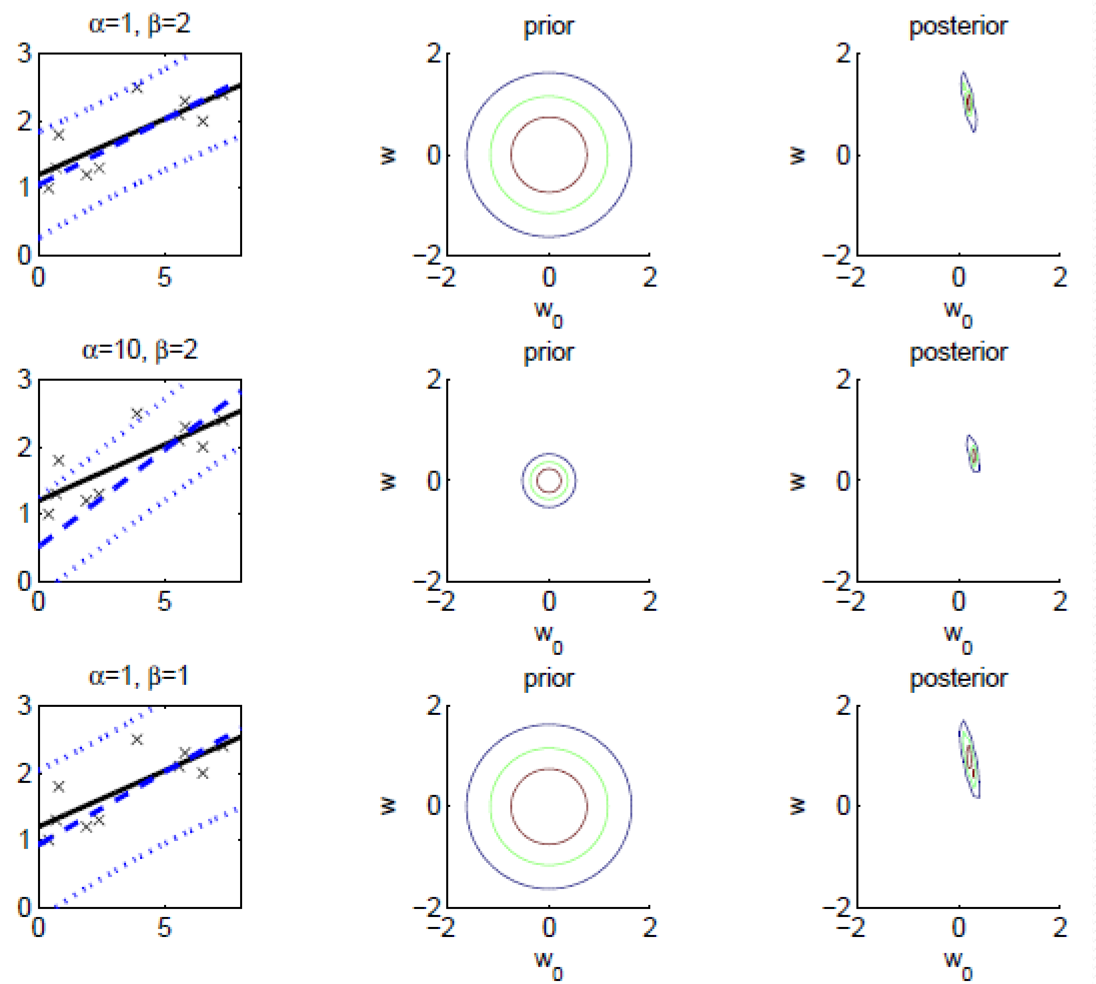
$$L(\mathbf{r} | \mathbf{X}, \mathbf{w}, \beta) = \log \prod_t p(r^t | \mathbf{x}^t, \mathbf{w}, \beta)$$

$$= -N \log\left(\sqrt{2\pi}\right) + N \log \beta - \frac{\beta}{2} \sum_t \left(r^t - \mathbf{w}^T \mathbf{x}^t\right)$$

- ML solution $\mathbf{w}_{ML} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{r}$
- Gaussian conjugate prior: $p(\mathbf{w}) \sim N(0, 1/\alpha)$
- Posterior: $p(\mathbf{w} | \mathbf{X}) \sim N(\boldsymbol{\mu}_N, \Sigma_N)$ where

$$\boldsymbol{\mu}_N = \beta \boldsymbol{\Sigma}_N \mathbf{X}^T \mathbf{r}$$

$$\boldsymbol{\Sigma}_N = (\alpha \mathbf{I} + \beta \mathbf{X}^T \mathbf{X})^{-1}$$


α=1, β=2
prior
posterior
α=10, β=2
prior
posterior
α=1, β=1
prior
posterior
w
w0

Basis/Kernel Functions

- For new $\boldsymbol{x}'$, the estimate r' is calculated as

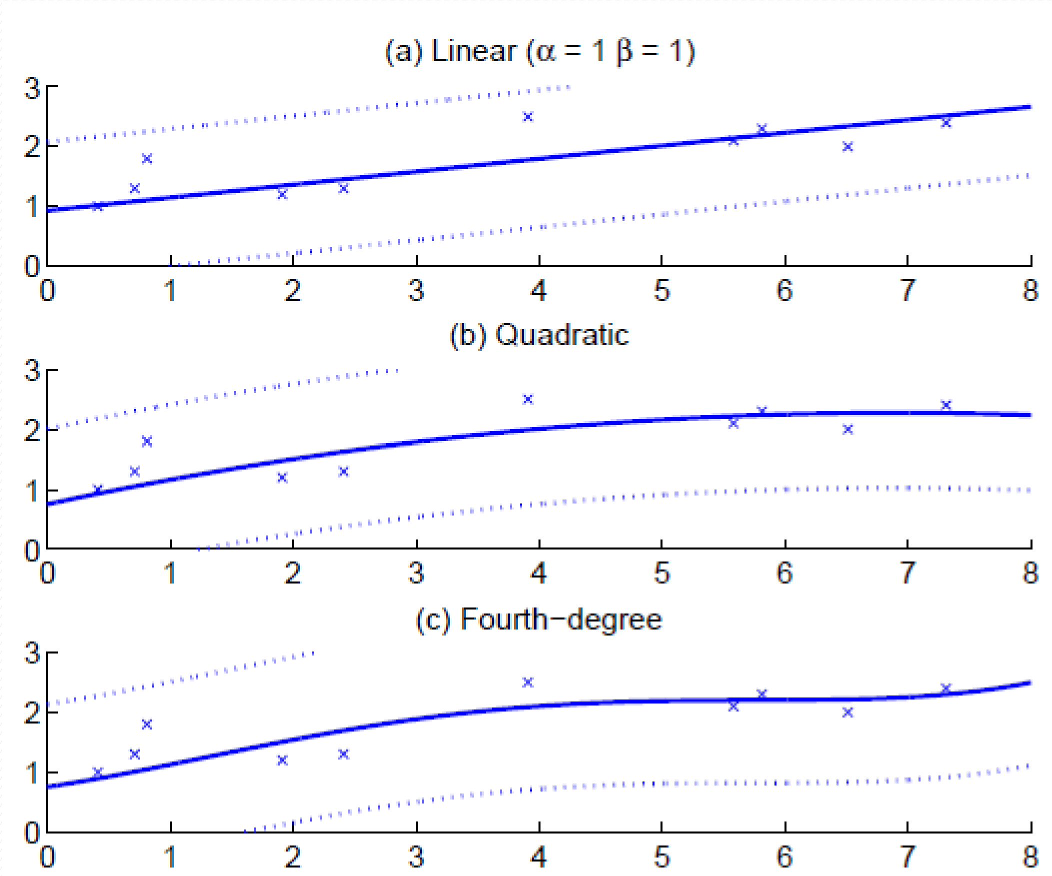
$$
\begin{aligned}
r' &= (\mathbf{x}')^T \\
&= \beta(\mathbf{x}')^T \mathbf{\Sigma}_N \mathbf{X}^T \mathbf{r} \\
&= \sum_t \beta(\mathbf{x}')^T \mathbf{\Sigma}_N \mathbf{x}^t r^t
\end{aligned}
$$

Dual representation

- Linear kernel $$r' = \sum_t \beta(\mathbf{x}')^T \mathbf{\Sigma}_N \mathbf{x}^t r^t \sum_t \beta K(\mathbf{x}', \mathbf{x}^t) r^t$$

- For any other $\phi(\mathbf{x})$, we can write $K(\mathbf{x}',\mathbf{x})=\phi(\mathbf{x}')^T\phi(\mathbf{x})$

Kernel Functions



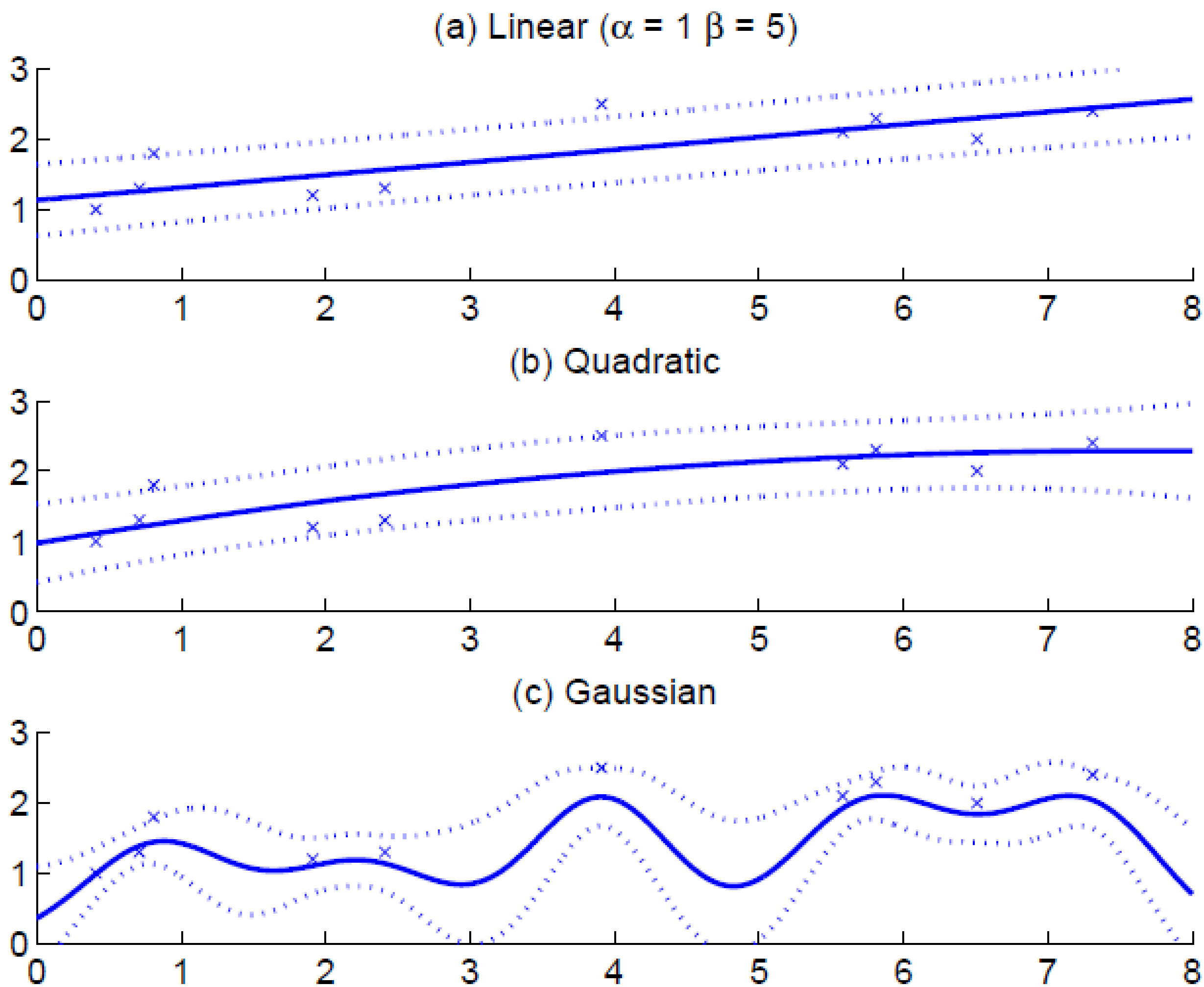
Gaussian Processes

- Assume Gaussian prior $p(\mathbf{w})\sim N(0,1/\alpha)$
- $\mathbf{y}=\mathbf{X}\mathbf{w}$, where $E[\mathbf{y}]=0$ and $\mathrm{Cov}(\mathbf{y})=\mathbf{K}$ with $\mathbf{K}_{ij}= (\mathbf{x}^i)^T\mathbf{x}^i$
- **K** is the covariance function, here linear
- With basis function $\phi(\mathbf{x})$, $\mathbf{K}_{ij}= (\phi(\mathbf{x}^i))^T\phi(\mathbf{x}^i)$
- $r\sim N_N(\mathbf{0},C_N)$ where $C_N= (1/\beta)\mathbf{I}+\mathbf{K}$
- With new $\mathbf{x}'$ added as $\mathbf{x}_{N+1}$, $r_{N+1}\sim N_{N+1}(0,C_{N+1})$

$$\mathbf{C}_{N+1}=\begin{bmatrix}\mathbf{C}_N & \mathbf{k}\\ \mathbf{k} & c\end{bmatrix}$$

where $\mathbf{k} = [K(\mathbf{x}',\mathbf{x}^t)_t]^T$ and $c=K(\mathbf{x}',\mathbf{x}')+1/\beta$.

$p(r'|\mathbf{x}',\mathbf{X},\mathbf{r})\sim N(\mathbf{k}^T\mathbf{C}_{N\text{-}1}\mathbf{r},c\text{-}\mathbf{k}^T\mathbf{C}_{N\text{-}1}\mathbf{k})$


(a) Linear (α = 1 β = 5)
(b) Quadratic
(c) Gaussian
0
1
2
3
4
5
6
7
8