Slides modified from:

PATTERN RECOGNITION

AND MACHINE LEARNING

CHRISTOPHER M. BISHOP

Predictive Distribution (1)

Predict t for new values of x by integrating over w:

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)\,\mathrm{d}\mathbf{w}$$
$$= \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

where

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}).$$

$$\mathbf{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}.$$

The Evidence Approximation (1)

The fully Bayesian predictive distribution is given by

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})\,\mathrm{d}\mathbf{w}\,\mathrm{d}\alpha\,\mathrm{d}\beta$$

but this integral is intractable. Approximate with

$$p(t|\mathbf{t}) \simeq p\left(t|\mathbf{t},\widehat{\alpha},\widehat{\beta}\right) = \int p\left(t|\mathbf{w},\widehat{\beta}\right)p\left(\mathbf{w}|\mathbf{t},\widehat{\alpha},\widehat{\beta}\right)\,\mathrm{d}\mathbf{w}$$

where $\left(\widehat{\alpha},\widehat{\beta}\right)$ is the mode of $p(\alpha,\beta|\mathbf{t})$, which is assumed to be sharply peaked; a.k.a. *empirical Bayes, type II* or *generalized maximum likelihood,* or *evidence approximation*.

The Evidence Approximation (2)

From Bayes' theorem we have

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta)$$

and if we assume $p(\alpha,\beta)$ to be flat we see that

$$\begin{aligned} p(\alpha, \beta|\mathbf{t}) &\propto p(\mathbf{t}|\alpha, \beta) \\ &= \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}. \end{aligned}$$

The Evidence Approximation (3)

Cont.:

$$\begin{aligned} p(\alpha, \beta|\mathbf{t}) &\propto p(\mathbf{t}|\alpha, \beta) \\ &= \int p(\mathbf{t}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}. \end{aligned}$$

Evidence function:

$$p(\mathbf{t}|\alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\left\{-E(\mathbf{w})\right\}\,\mathrm{d}\mathbf{w}$$

with

$$\begin{aligned} E(\mathbf{w}) &= \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \\ &= \frac{\beta}{2}\left\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right\|^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \end{aligned}$$

The Evidence Approximation (4)

Cont.:

$$\begin{aligned} E(\mathbf{w}) &= \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \\ &= \frac{\beta}{2}\left\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\right\|^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \end{aligned}$$

Completing the square over w:

$$E(\mathbf{w}) = E(\mathbf{m}_N) + \frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w} - \mathbf{m}_N)$$

with

$$E(\mathbf{m}_N) = \frac{\beta}{2}\left\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{m}_N\right\|^2 + \frac{\alpha}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N$$

$$\mathbf{A} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

$$\mathbf{A} = \mathbf{S}_N^{-1}$$

$$\mathbf{m}_N = \beta\mathbf{A}^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}$$

The Evidence Approximation (5)

Evaluate integral over w

$$\int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w}$$

$$= \exp\{-E(\mathbf{m}_N)\}\int \exp\left\{-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{A}(\mathbf{w}-\mathbf{m}_N)\right\}\mathrm{d}\mathbf{w}$$

$$= \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2}.$$

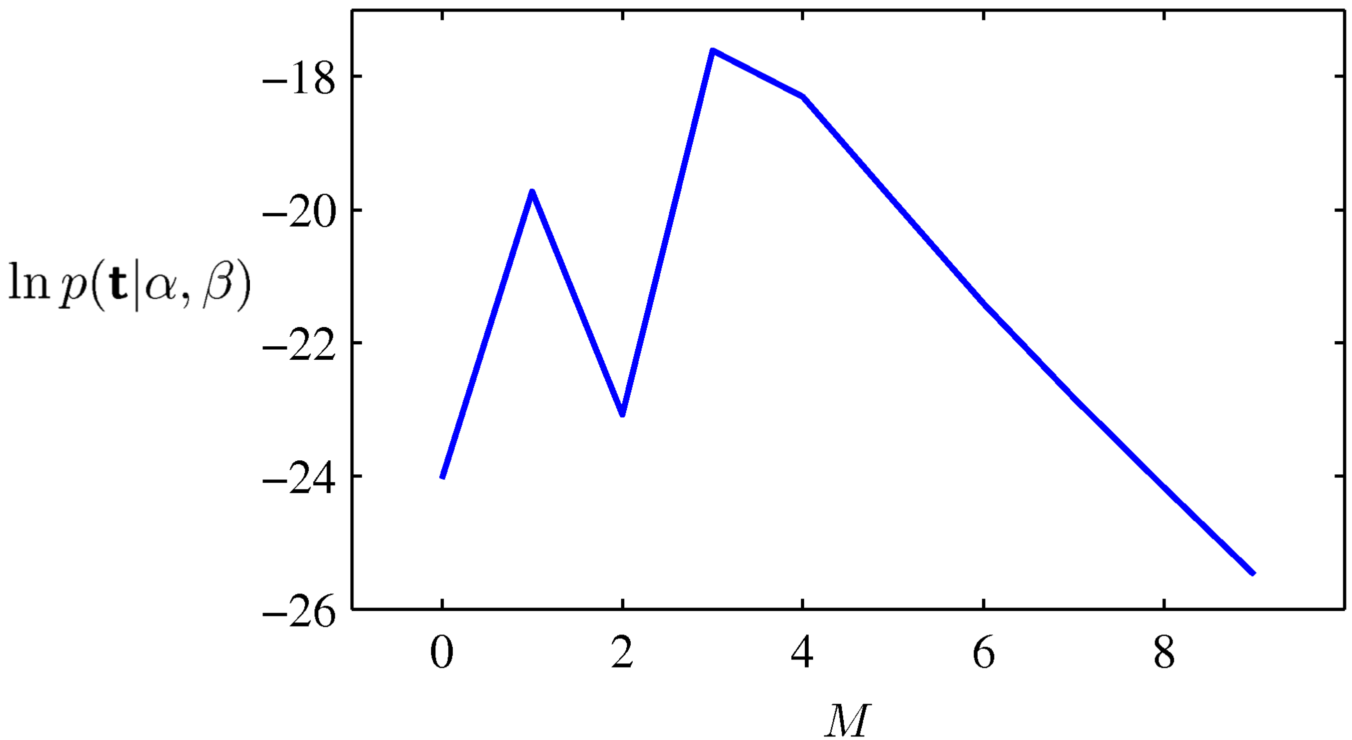
Thus, log of marginal likelihood (evidence function):

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) + \frac{1}{2}\ln|\mathbf{S}_N| - \frac{N}{2}\ln(2\pi).$$

The Evidence Approximation (6)

Example: sinusoidal data, M th degree polynomial,

$$\alpha = 5 \times 10^{-3}$$



Maximizing the Evidence Function (1)

To maximise $\ln p(\mathbf{t}|\alpha, \beta)$ w.r.t. α and β, we define the eigenvector equation

$$\left(\beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right) \mathbf{u}_i = \lambda_i \mathbf{u}_i.$$

Thus

$$\mathbf{A} = \mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}$$

has eigenvalues $\lambda_i + \alpha$.

Maximizing the Evidence Function (2)

Derivative of $\ln|\mathbf{A}|$ with respect to α

$$\frac{d}{d\alpha}\ln|\mathbf{A}| = \frac{d}{d\alpha}\ln\prod_i(\lambda_i+\alpha) = \frac{d}{d\alpha}\sum_i\ln(\lambda_i+\alpha) = \sum_i\frac{1}{\lambda_i+\alpha}$$

Stationary points of log marginal likelihood

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N - \frac{1}{2}\sum_i\frac{1}{\lambda_i+\alpha}$$

Thus

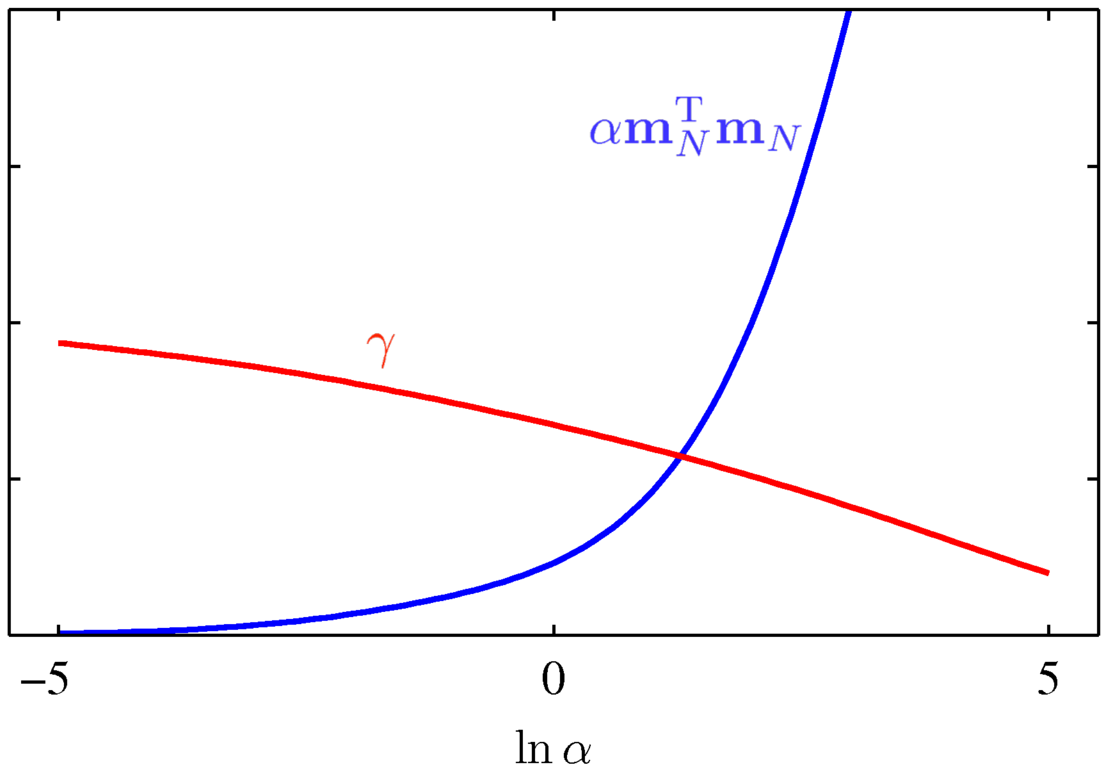
$$\alpha\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N = M - \alpha\sum_i\frac{1}{\lambda_i+\alpha} = \gamma$$

and therefore

$$\gamma = \sum_i\frac{\lambda_i}{\alpha+\lambda_i}$$

Maximizing the Evidence Function (3)

Example: sinusoidal data, 9 Gaussian basis functions, $\beta = 11.1$.



Maximizing the Evidence Function (4)

Thus differentiating $\ln p(\mathbf{t}|\alpha, \beta)$ w.r.t. α and β, and set the results to zero, to get

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N}$$

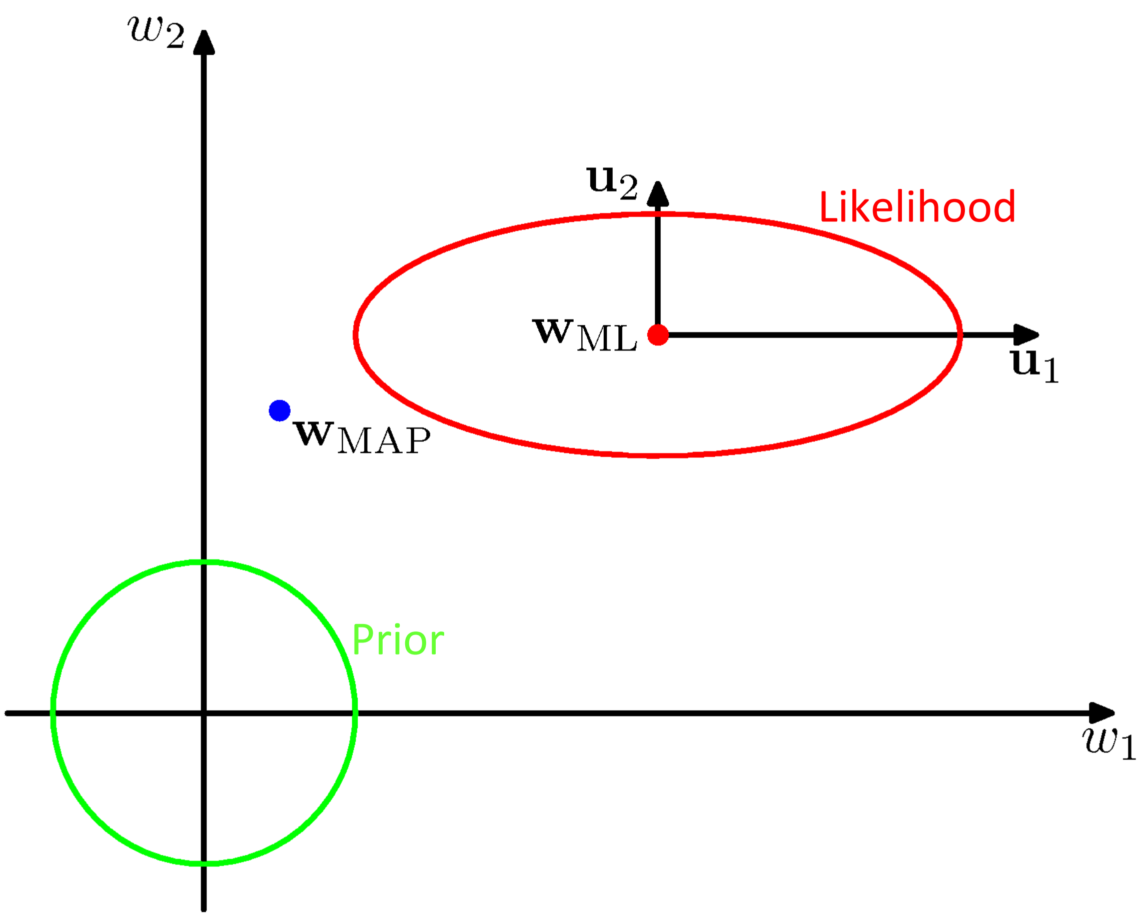
$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2$$

where

$$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$$

Note γ depends on both α and β.

Effective Number of Parameters (1)



$\lambda_1 \ll \alpha$

w_1 is not well determined by the likelihood

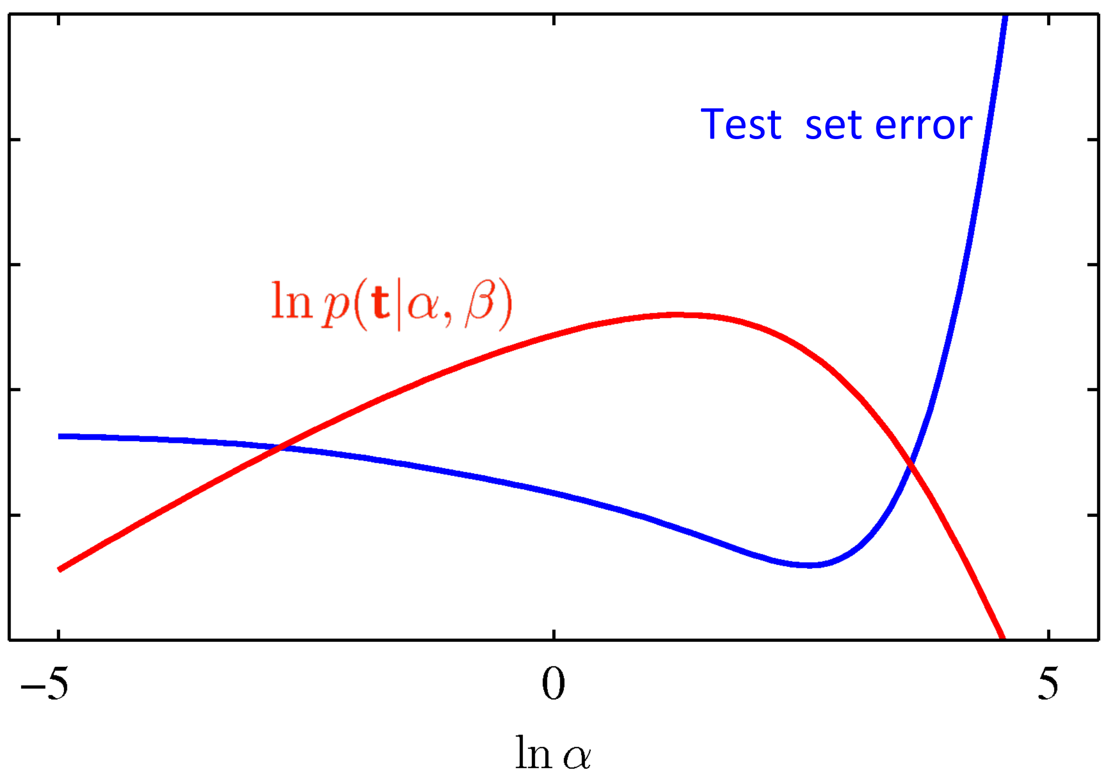
$\lambda_2 \gg \alpha$

w_2 is well determined by the likelihood

γ is the number of well determined parameters

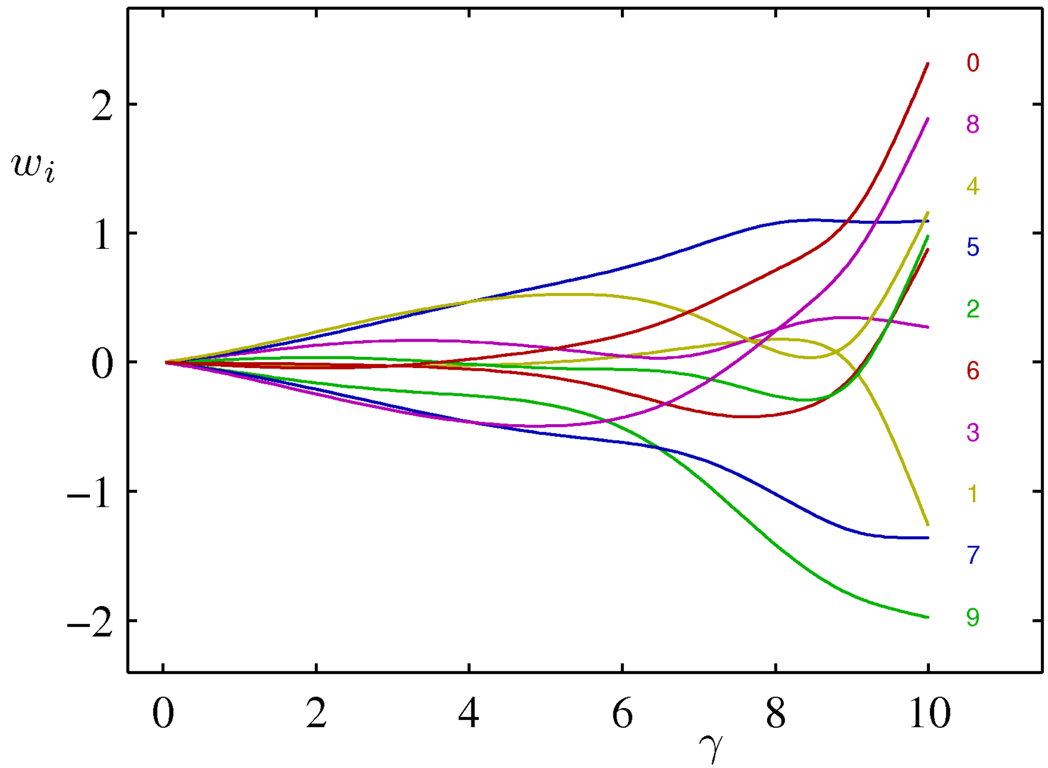
Effective Number of Parameters (3)

Example: sinusoidal data, 9 Gaussian basis functions, $\beta = 11.1$.



Effective Number of Parameters (4)

Example: sinusoidal data, 9 Gaussian basis functions, $\beta = 11.1$.



Effective Number of Parameters (5)

In the limit $N \gg M$, $\gamma = M$ and we can consider using the easy-to-compute approximation

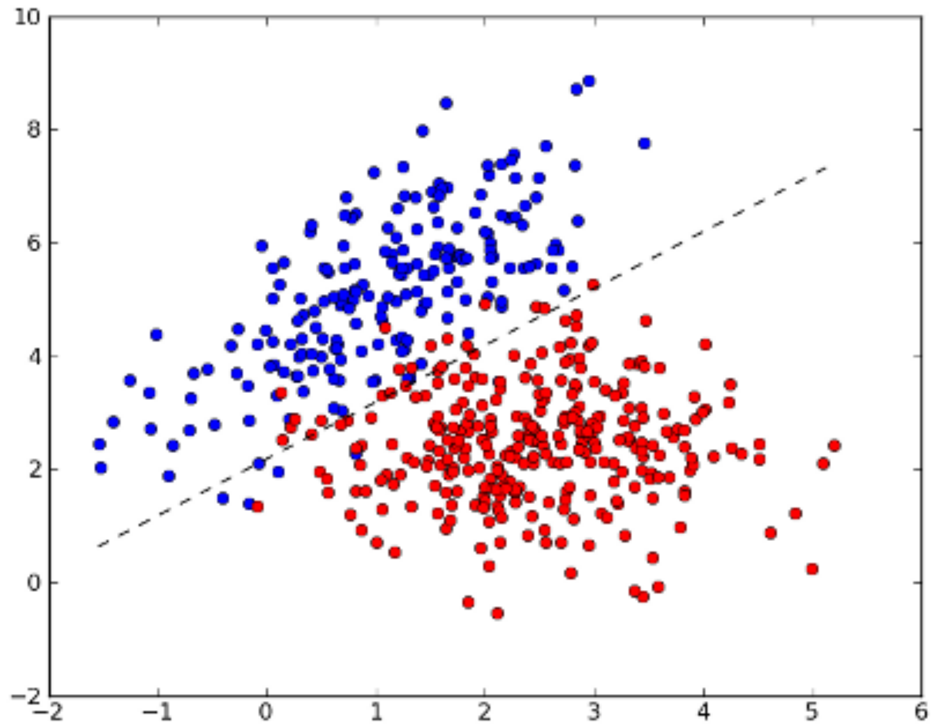
$$\alpha = \frac{M}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N}$$

$$\frac{1}{\beta} = \frac{1}{N}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2.$$

Limitations of Fixed Basis Functions

- Class of nonlinearities may be insufficient
- M basis function along each dimension of a D-dimensional input space requires M^D basis functions: the curse of dimensionality.
- Choosing basis functions using the training data.

Classification



Linear models for classification

Assign input vector **x** to one of k discrete classes C_k , k=1,…,K.

D-dimensional input space

Decision boundary/surface: (D-1)-dimensional hyperplane

Regression vs. Classification

Regression:

$$x \in [-\infty, \infty], t \in [-\infty, \infty]$$

Classification (two classes):

$$x \in [-\infty, \infty], t \in \{0, 1\}$$

Regression vs. Classification

Linear regression model prediction (y real)

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

Classification: y in range (0,1) (posterior probabilities)

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right)$$

f: Activation function (nonlinear)

Decision surface: $\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0 = \text{constant}$

(Generalized linear models)

Binary Variables (1)

Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

Bernoulli Distribution

$$\begin{aligned} \mathrm{Bern}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

Binary Variables (2)

N coin flips:

$$p(m \text{ heads}|N, \mu)$$

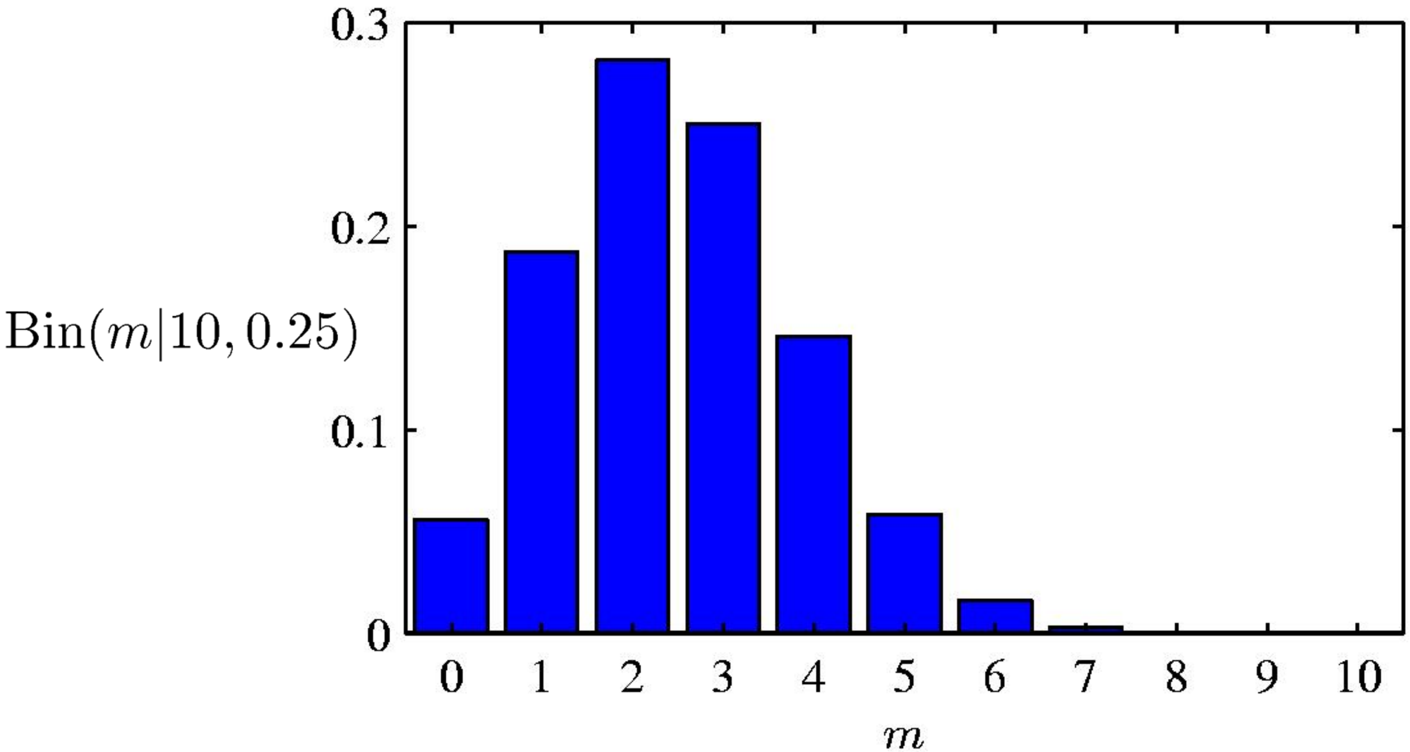
Binomial Distribution

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N,\mu) = N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \, \mathrm{Bin}(m|N,\mu) = N\mu(1-\mu)$$

Binomial Distribution



Parameter Estimation (1)

ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n)\ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$$

Parameter Estimation (2)

Example: $$\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$$

Prediction: *all* future tosses will land heads up

Overfitting to D

Decision Theory

Inference step

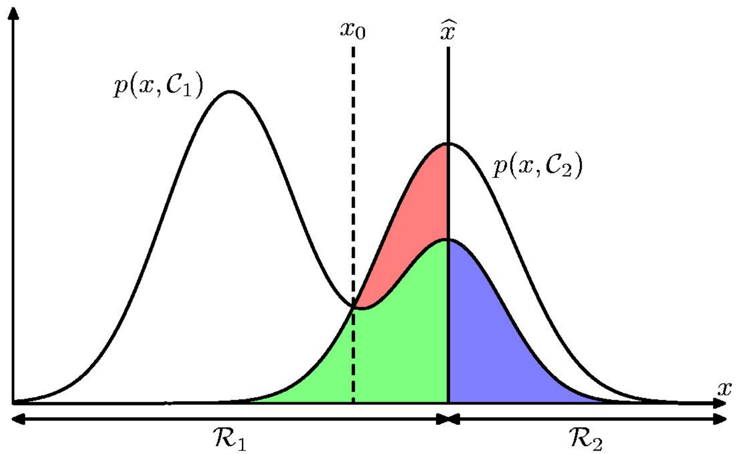
Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$.

Decision step

For given x, determine optimal t.

Minimum Misclassification Rate

$$p(\text{mistake}) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1)$$
$$= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}.$$



We are free to choose the decision rule that assigns each point x to one of the two classes. This defines the decision regions Rk.

To minimize integrand: $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$ must be small

Assign x to class for which the posterior $p(\mathcal{C}_k|\mathbf{x})$ is larger!