CSCI 5521: Machine Learning Fundamentals

Parametric Models (Chpt 3-5)

Rui Kuang

Department of Computer Science and Engineering

University of Minnesota

Probabilistic Perspective

- We have seen classification models. Classification decision is deterministic

$$h(\mathbf{x}) = \begin{cases} 1 \text{ if } h \text{ says } \mathbf{x} \text{ is positive} \\ 0 \text{ if } h \text{ says } \mathbf{x} \text{ is negative} \end{cases}$$

- What if we have cases with some uncertainty?
- Estimation of

$$p(C = 0 \mid \mathbf{x}) \text{ and } P(C = 1 \mid \mathbf{x})$$

Classification

- Credit scoring: Inputs are income and savings. Output is low-risk vs high-risk
- Input: $\boldsymbol{x} = [x_1, x_2]^T$,Output: C is in {0,1}

- Prediction:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C = 1 | x_1, x_2) > P(C = 0 | x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

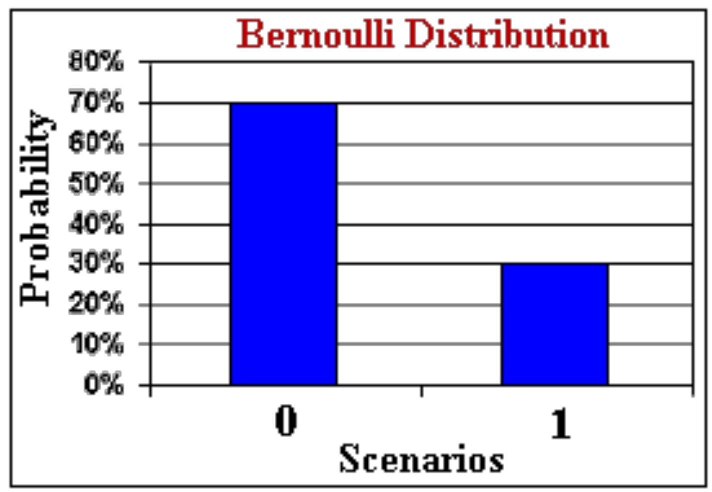
Coin Toss

- Random var $X \in \{1,0\}$

  Bernoulli: $P\{X=1\} = p_o^X (1-p_o)^{(1-X)}$
- Sample: $\mathbf{X} = \{x^t\}_{t=1}^N = \{0,1,\ldots,1,\ldots\}$
- Prediction of next toss (no input):

  Heads if $p_o > ½$,

  Tails otherwise



Parametric Estimation

- $\mathcal{X} = \{ x^t \}$ where $x^t \sim p(x)$
- Parametric estimation:

  Assume a form for $p(x \mid \theta)$ and estimate θ, its parameters, using X

  e.g., N (μ, σ^2) where $\theta = \{ \mu, \sigma^2 \}$

  $$P(x) = p^x (1 - p)^{(1 - x)}$$

Maximum Likelihood Estimation

- Likelihood of θ given the sample $\mathcal{X}$

  $l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_t p(x^t|\theta)$

- Log likelihood

  $\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta)$

- Maximum likelihood estimator (MLE)

  $\theta^* = \text{argmax}_\theta \mathcal{L}(\theta|\mathcal{X})$

Examples: Bernoulli

- Bernoulli: Two states, x in {0,1}

$P(x) = p^x (1 - p)^{(1 - x)}$

$\mathcal{L}(p|\mathcal{X}) = \log \prod_t p^{x^t} (1 - p)^{(1 - x^t)}$

$$L(p \mid X) = \sum_t x^t \log p + \left(N - \sum_t x^t\right) \log(1 - p)$$

$$\frac{\partial L(p \mid X)}{p} = \sum_t x^t / p - \left(N - \sum_t x^t\right) / (1 - p) = 0$$

MLE: $p = \sum_t x^t / N$

Examples: Multinomial (Categorical)

- Multinomial: K>2 states, x_i in {0,1}

$$P(x_1,x_2,\ldots,x_K) = \prod_{i=1\ldots K} p_i^{x_i}$$

$$\mathcal{L}(p_1,p_2,\ldots,p_K|\mathcal{X}) = \log \prod_{t=1\ldots N} \prod_{i=1\ldots K} p_i^{x_i^t}$$

$$L(p_1 \ldots p_K \mid \chi) = \sum_t \sum_i x_i^t \log p_i \qquad \text{with } \sum_i p_i = 1.$$

$$\frac{\partial\left(\sum_t \sum_i x_i^t \log p_i - \alpha\left(\sum_i p_i - 1\right)\right)}{\partial p_i} = 0$$

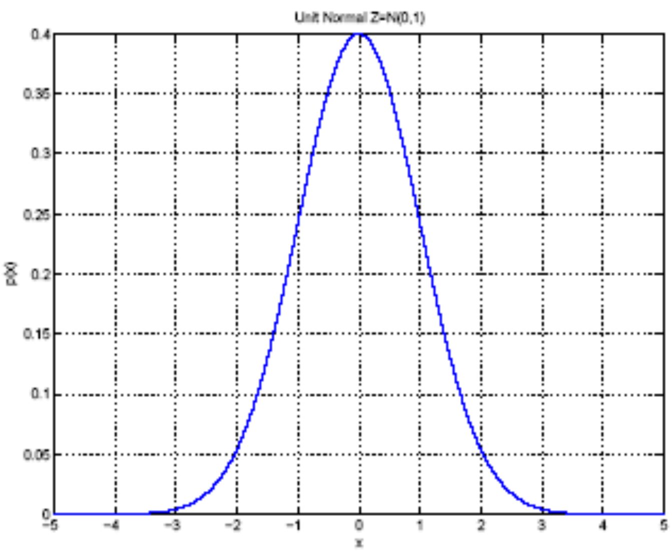
$$\sum_t x_i^t / p_i - \alpha = 0$$

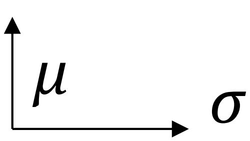
MLE: $p_i = \sum_t x_i^t / N$

Gaussian (Normal) Distribution

- $p(x) = \mathcal{N}(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$





$$L(\mu,\sigma \mid X) = \sum_t \log \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x^t-\mu)^2}{2\sigma^2}\right]$$

$$L(\mu,\sigma \mid X) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t-\mu)^2}{2\sigma^2}$$

$$\frac{\partial L(\mu,\sigma \mid X)}{\partial \mu} = \sum_t (x^t-\mu) = 0 \Rightarrow \mu = \frac{\sum_t x^t}{N}$$

$$\frac{\partial L(\mu,\sigma \mid X)}{\partial \sigma} = \frac{N}{\sigma} - \frac{\sum_t (x^t-\mu)^2}{\sigma^3} = 0$$

$$\sigma^2 = \frac{\sum_t (x^t-\mu)^2}{N}$$

Bayes' Rule

- How to get $P(C|x)$?

posterior

prior

likelihood

$$P(C \mid \mathbf{x}) = \frac{P(C)\,p(\mathbf{x} \mid C)}{p(\mathbf{x})}$$

evidence

$$P(C=0)+P(C=1)=1$$

$$p(\mathbf{x}) = p(\mathbf{x}|C=1)P(C=1)+p(\mathbf{x}|C=0)P(C=0)$$

$$p(C=0|\mathbf{x})+P(C=1|\mathbf{x})=1$$

Bayes' Rule: General Formula

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i)P(C_i)}{p(\mathbf{x})}$$

$$= \frac{p(\mathbf{x} \mid C_i)P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x} \mid C_k)P(C_k)}$$

$$P(C_i) \geq 0 \text{ and } \sum_{i=1}^{K} P(C_i) = 1$$

$$\text{choose } C_i \text{ if } P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$$

Bayes' Rule Example

prior *likelihood*

$$P(C \mid \mathbf{x}) = \frac{P(C)\,p(\mathbf{x} \mid C)}{p(\mathbf{x})}$$

posterior *evidence*

P(x=high,med,low|acc) = 0.6, 0.4, 0

P(x=high,med,low|unacc) = 0, 1/6, 5/6

What if we don't use Bayes' rule?

Safty (x)	Rating (C)
'high'	'acc'
'low'	'unacc'
'med'	'acc'
'high'	'acc'
'low'	'unacc'
'med'	'acc'
'high'	'acc'
'low'	'unacc'
'med'	'unacc'
'high'	'acc'
'low'	'unacc'
'med'	'acc'
'high'	'acc'
'low'	'unacc'
'med'	'acc'
'high'	'acc'

Parametric Classification

- Discriminant function

$$g_i(x) = p(x \mid C_i)P(C_i)$$

or

$$g_i(x) = \log p(x \mid C_i) + \log P(C_i)$$

- Gaussians:

$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

- Given the sample

$$\mathcal{X} = \{x^t, r^t\}_{t=1}^N \qquad x \in \Re \qquad r_i^t = \begin{cases} 1 \text{ if } x^t \in C_i \\ 0 \text{ if } x^t \in C_j, j \neq i \end{cases}$$

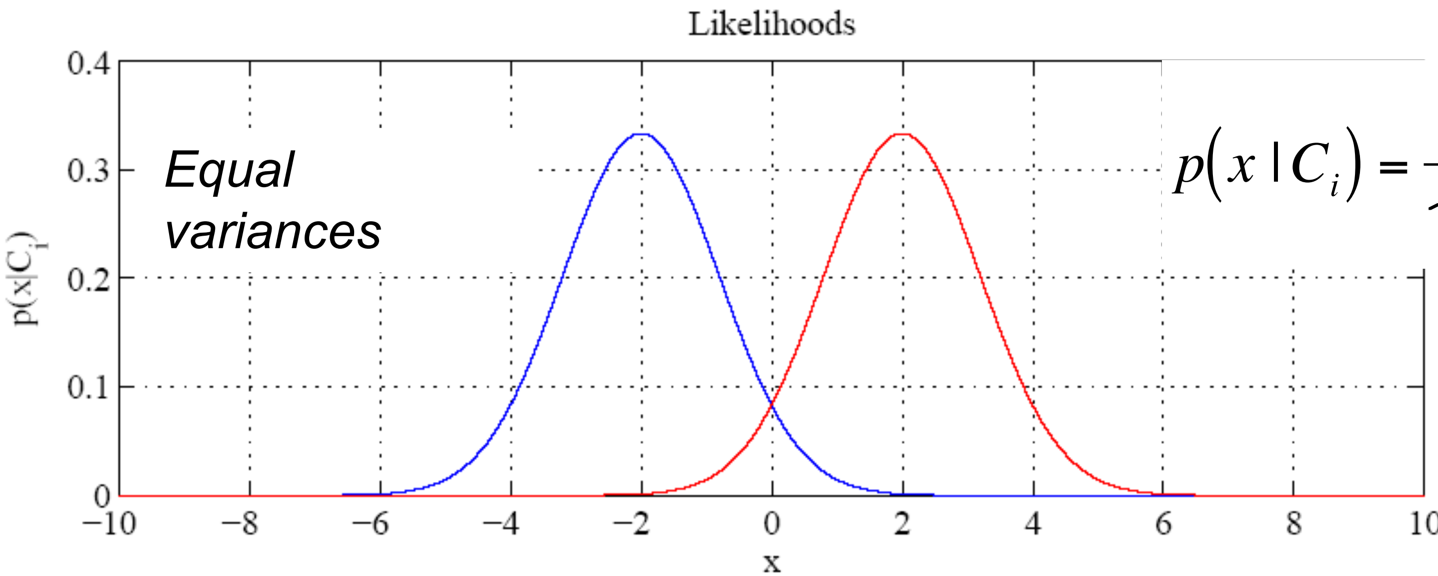
- ML estimates are

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

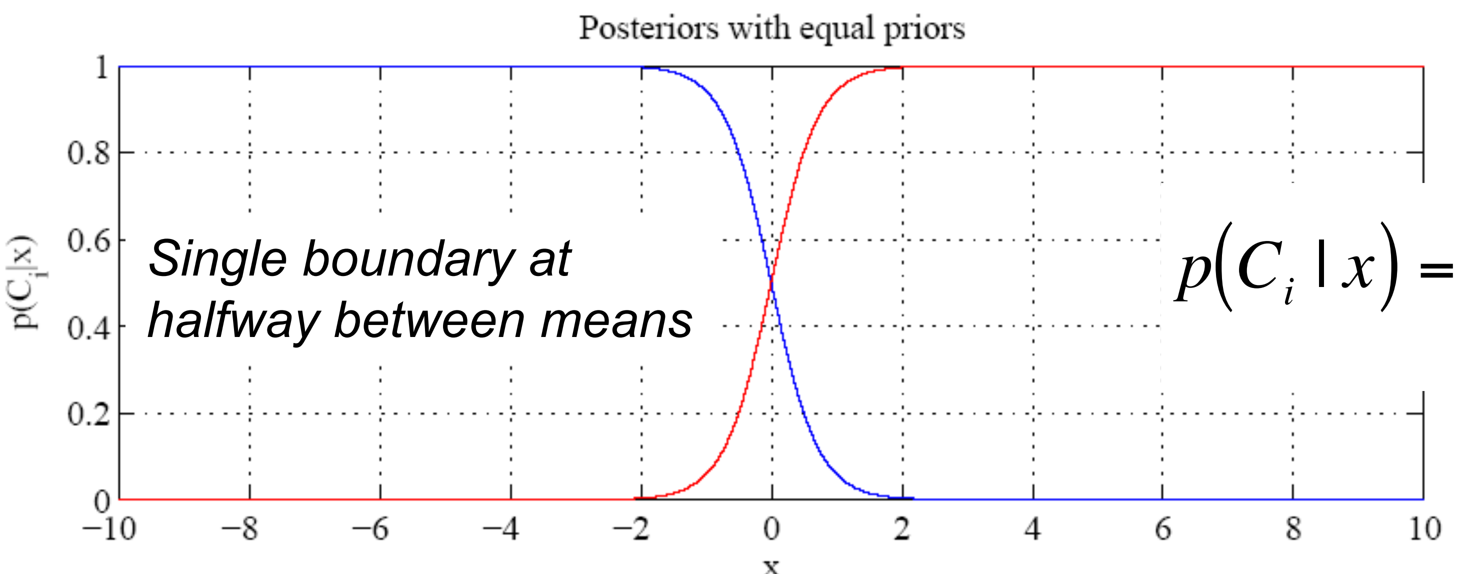
- Discriminant becomes

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x - m_i)^2}{2{s_i}^2} + \log \hat{P}(C_i)$$

Discriminant function: $g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x-m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$

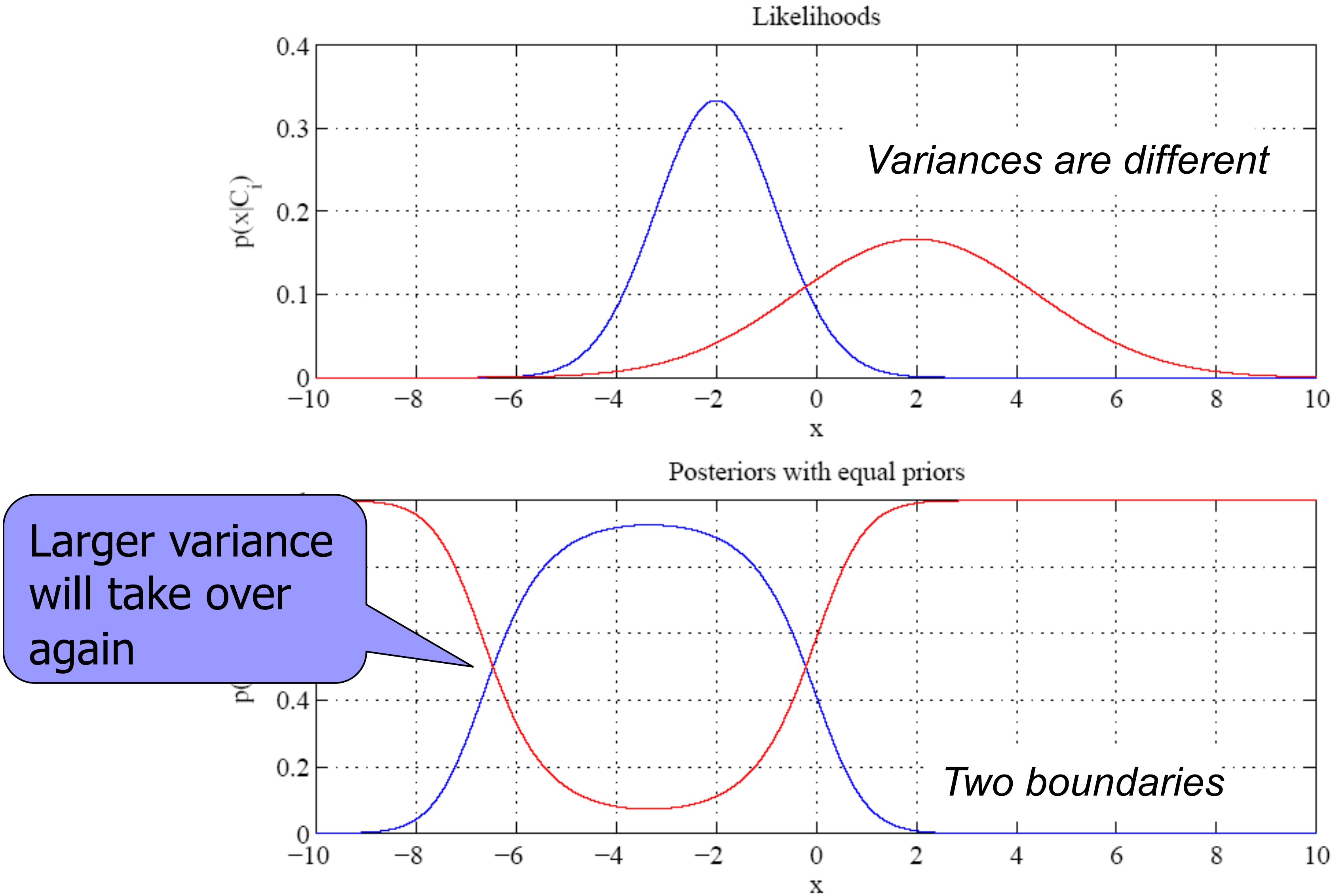


$$p(x \mid C_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$



$$p(C_i \mid x) = \frac{p(x \mid C_i)\,p(C_i)}{\sum_k p(x \mid C_k)\,p(C_k)}$$

E. Alpaydin, Introduction to Machine Learning



Likelihood-based approach: estimate densities separately.

Multivariate Data

- Multiple measurements (sensors)
- *d* inputs/features/attributes: *d*-variate
- *N* instances/observations/examples

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

Discrete Features

- Binary features: $p_{ij} \equiv p\left(x_j = 1 \mid C_i\right)$

  if x_j are independent (Naive Bayes')

$$p\left(x \mid C_i\right) = \prod_{j=1}^{d} p_{ij}^{x_j}\left(1 - p_{ij}\right)^{\left(1 - x_j\right)}$$

Example: Text classification. Each dimension is a word, set to 1 if word in the text

			x^1	x^2	x^3	x^4
Dim1:	"the"	=	1	0	1	1
Dim2:	"hello"	=	0	1	0	1
Dim3:	"and"	=	1	1	0	1
Dim4:	"happy"	=	1	0	0	1

Discrete Features

- Likelihood of Naive Bayes': the discriminant is linear

$$g_i(\mathbf{x}) = \log p(\mathbf{x} \mid C_i) + \log P(C_i)$$

$$= \sum_j \left[x_j \log p_{ij} + (1 - x_j) \log (1 - p_{ij}) \right] + \log P(C_i)$$

Estimated parameters $\hat{p}_{ij} = \frac{\sum_t x_j^t r_i^t}{\sum_t r_i^t}$

Discrete Features

- Multinomial (1-of-n_j) features: $x_j = \{v_1, v_2, \ldots, v_{n_j}\}$

$$p_{ijk} \equiv p(z_{jk} = 1 \mid C_i) = p(x_j = v_k \mid C_i)$$

In class C_i, variable x_j is category v_k

if x_j are independent

$$p(\mathbf{x} \mid C_i) = \prod_{j=1}^{d} \prod_{k=1}^{n_j} p_{ijk}^{z_{jk}}$$

$$g_i(\mathbf{x}) = \sum_j \sum_k z_{jk} \log p_{ijk} + \log P(C_i)$$

$$\hat{p}_{ijk} = \frac{\sum_t z_{jk}^t r_i^t}{\sum_t r_i^t}$$

Multivariate Parameters

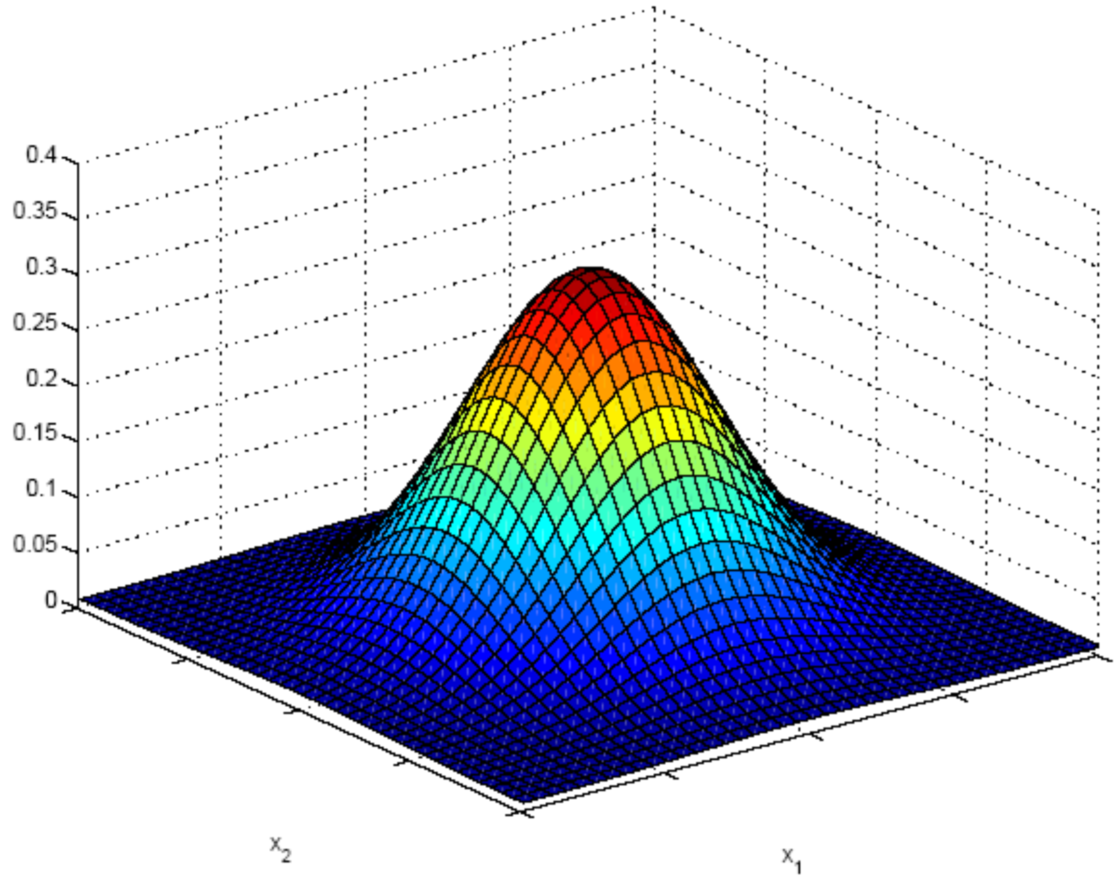
$$\text{Mean}: E[\mathbf{x}] = \mu = [\mu_1, \ldots, \mu_d]^T$$

$$\text{Covariance}: \sigma_{ij} \equiv \text{Cov}(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$$

$$\text{Correlation}: \ \text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$\Sigma \equiv \text{Cov}(\mathbf{X}) = E\left[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Multivariate Normal Distribution



$$\mathbf{x} \sim \mathcal{N}_d(\mu, \Sigma)$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right]$$

Independent Inputs: Naive Bayes

- If x_i are independent, offdiagonals of ∑ are 0, Mahalanobis distance reduces to weighted (by $1/\sigma_i$) Euclidean distance:

$$p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) = \frac{1}{(2\pi)^{d/2}\prod_{i=1}^{d}\sigma_i}\exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$

- If variances are also equal, reduces to Euclidean distance

Multivariate Normal Distribution

- Mahalanobis distance: $(\boldsymbol{x} - \boldsymbol{\mu})^T \sum^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ measures distance from x to μ by rotation and normalization with ∑
  - normalizes for difference in variances and correlations
  - Variable with larger varance receive less weight
  - Two highly correlated variables contribute less.

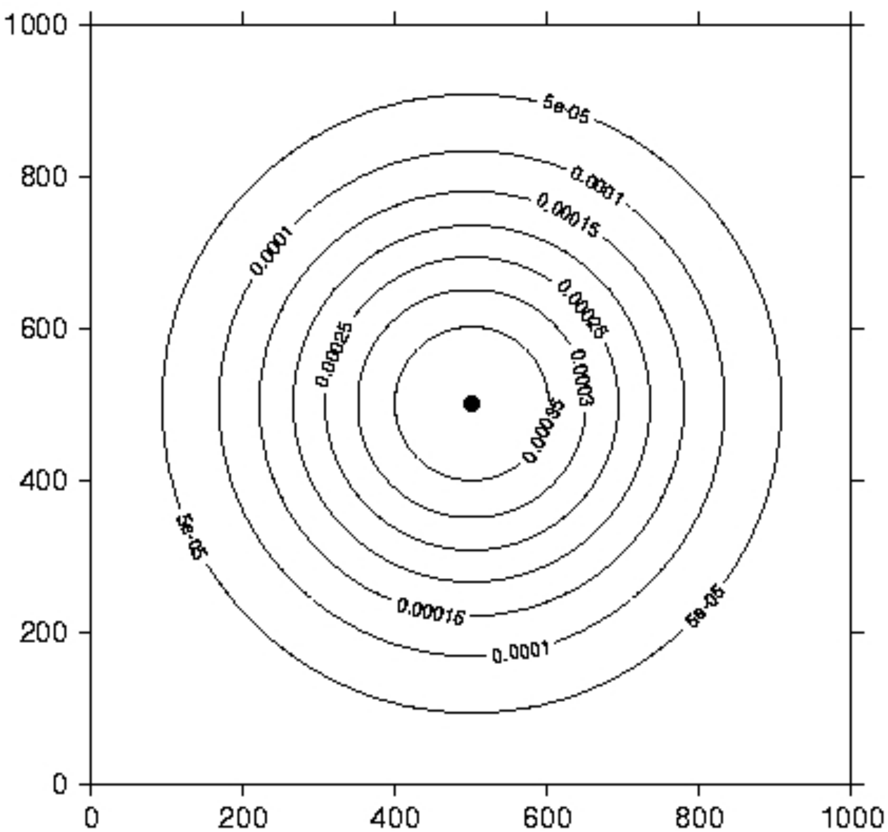
Multivariate Normal Distribution

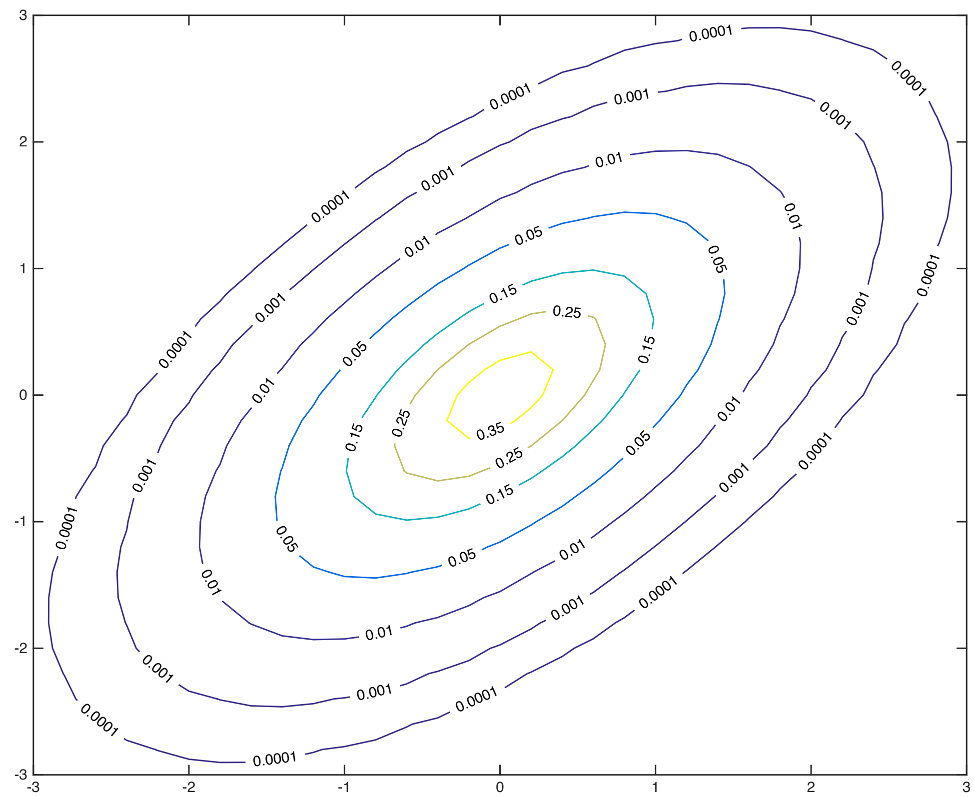
- Covariances are obtained by rotation matrix *P*

$$(x-\boldsymbol{\mu})^T\,\Sigma^{-1}(x-\boldsymbol{\mu})$$

$$= (x-\boldsymbol{\mu})^T\,P^T\begin{pmatrix} 1/{\sigma_1}^2 & & & & & \\ & 1/{\sigma_2}^2 & & & & \\ & & \cdots & & & \\ & & \cdots & & & \\ & & \cdots & & & \\ & & & & & 1/{\sigma_d}^2 \end{pmatrix}P(x-\boldsymbol{\mu})$$

Interpreting Probability Density Contour





All the points on a contour curve has the same Mahalanobis distance to the mean.

Multivariate Normal Distribution

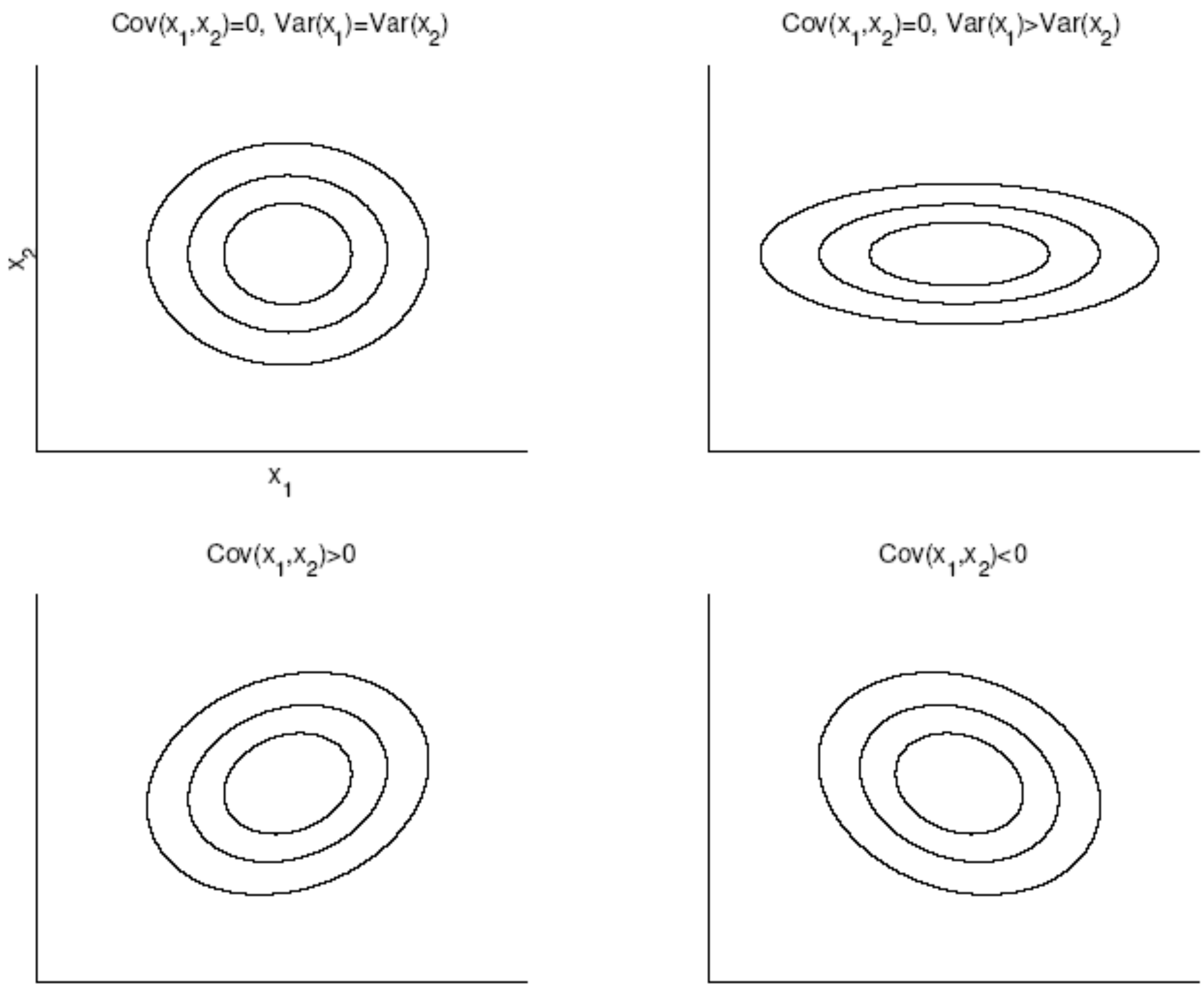
- Bivariate: $d = 2$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

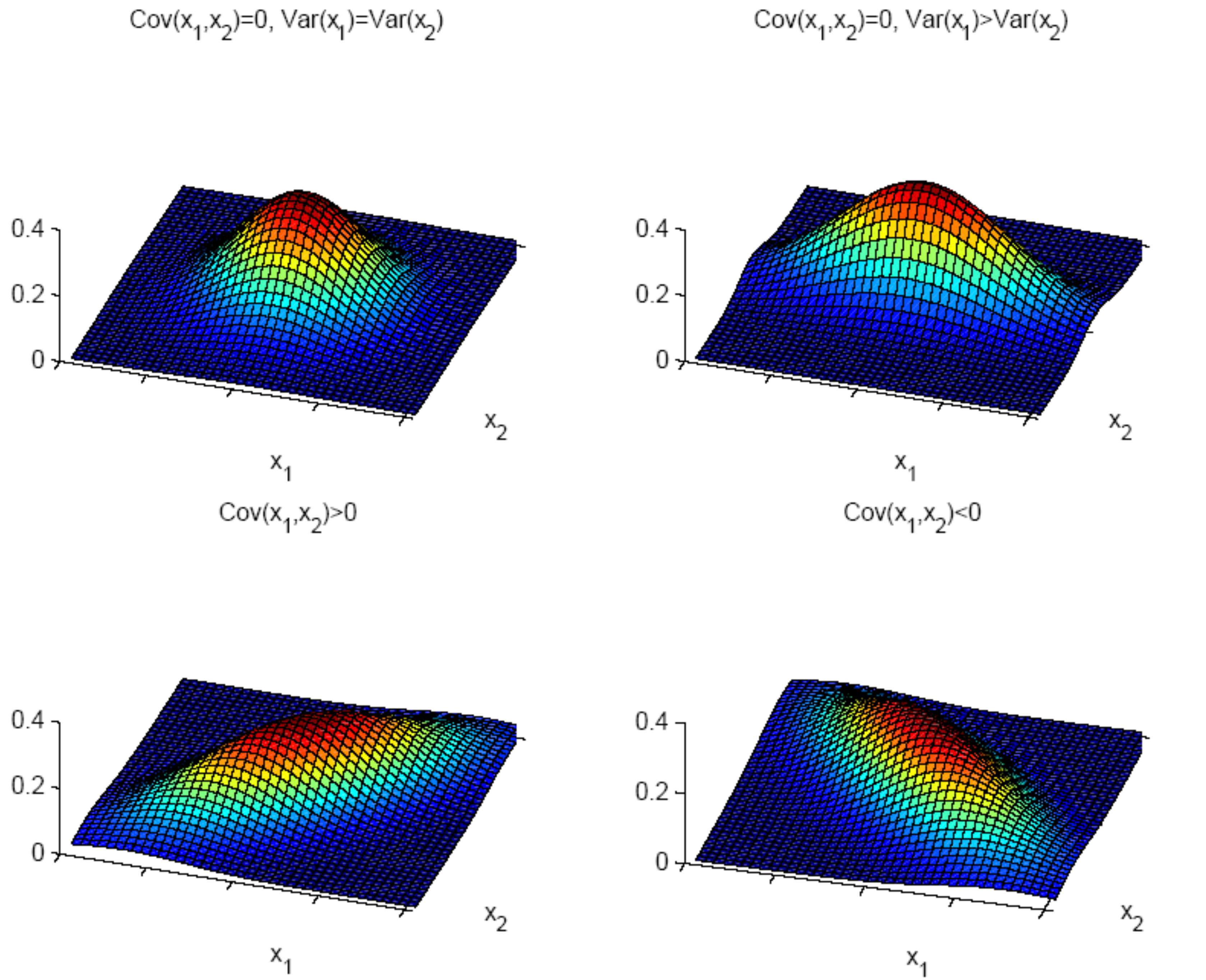
$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right)\right]$$

$$z_i = (x_i - \mu_i)/\sigma_i$$

Bivariate Normal



E. Alpaydin, Introduction to Machine Learning



E. Alpaydin, Introduction to Machine Learning

Max Likelihood for Multivariate Gaussian

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right]$$

$$L(\mu,\Sigma \mid \chi) = \sum_{t=1}^{N} -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{x}^t-\mu)^T \Sigma^{-1} (\mathbf{x}^t-\mu)$$

$$\text{max over } \mu : \frac{\partial L(\mu,\Sigma \mid \chi)}{\partial \mu} = \frac{\partial \sum_{t=1}^{N} -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{x}^t-\mu)^T \Sigma^{-1} (\mathbf{x}^t-\mu)}{\partial \mu}$$

$$\sum_{t=1}^{N} (\mathbf{x}^t-\mu)^T \Sigma^{-1} = 0 \qquad\qquad \mu = \frac{1}{N}\sum_{t=1}^{N} x^t$$

Max Likelihood for Multivariate Gaussian

$$\text{max over } \Sigma: \frac{\partial L(\mu,\Sigma\,|\,\chi)}{\partial \Sigma^{-1}} = \frac{\partial \sum_{t=1}^{N} -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(\mathbf{x}^t-\mu\right)^T \Sigma^{-1}\left(\mathbf{x}^t-\mu\right)}{\partial \Sigma^{-1}}$$

$$= \frac{\partial \sum_{t=1}^{N} -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2} trace[\left(\mathbf{x}^t-\mu\right)\left(\mathbf{x}^t-\mu\right)^T \Sigma^{-1}]}{\partial \Sigma^{-1}}$$

$$\frac{\partial \ln|A^{-1}|}{\partial A} = -(A^{-1})^T \qquad \frac{\partial trace[BA]}{\partial A} = B^T$$

$$= \frac{N}{2}\Sigma - \frac{1}{2}\sum_{t=1}^{N}\left(x^t-\mu\right)\left(x^t-\mu\right)^T$$

$$\Sigma = \frac{\sum_{t=1}^{N}\left(x^t-\mu\right)\left(x^t-\mu\right)^T}{N}$$

Parametric Classification

- If $p\ (\boldsymbol{x} \mid C_i) \sim \mathrm{N}\ (\boldsymbol{\mu}_i\ ,\ \Sigma_i\ )$

$$p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_i\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right]$$

- Discriminant functions

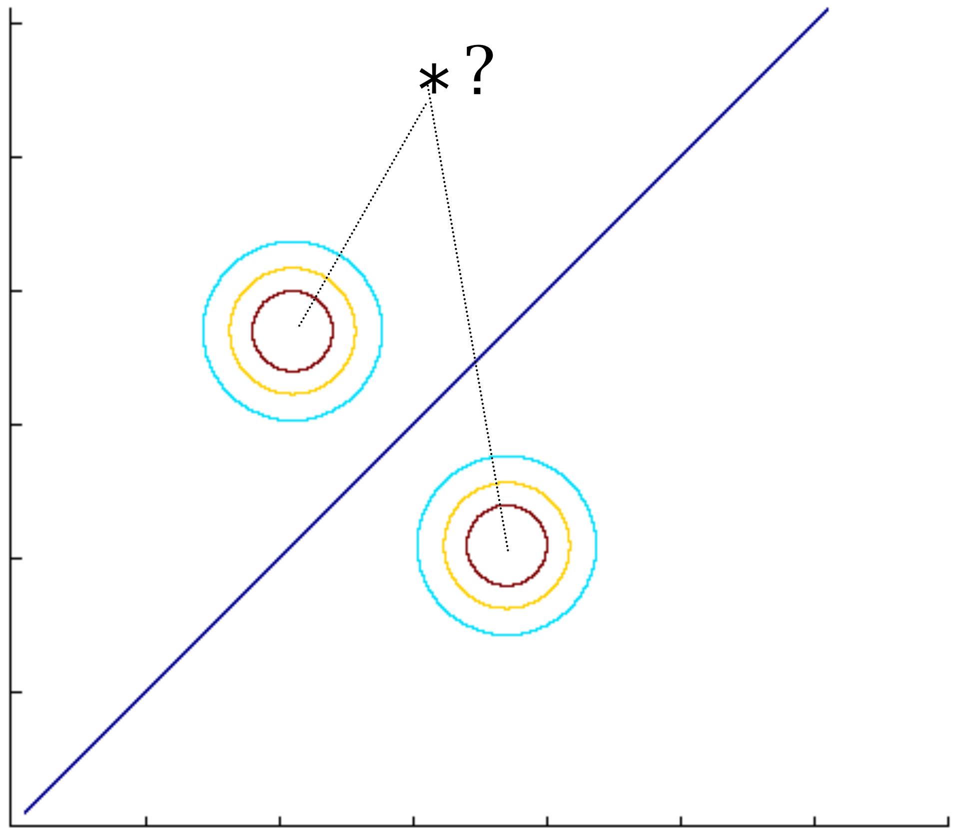
$$\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} \mid C_i) + \log P(C_i) \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log\left|\Sigma_i\right| - \frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \log P(C_i) \end{aligned}$$

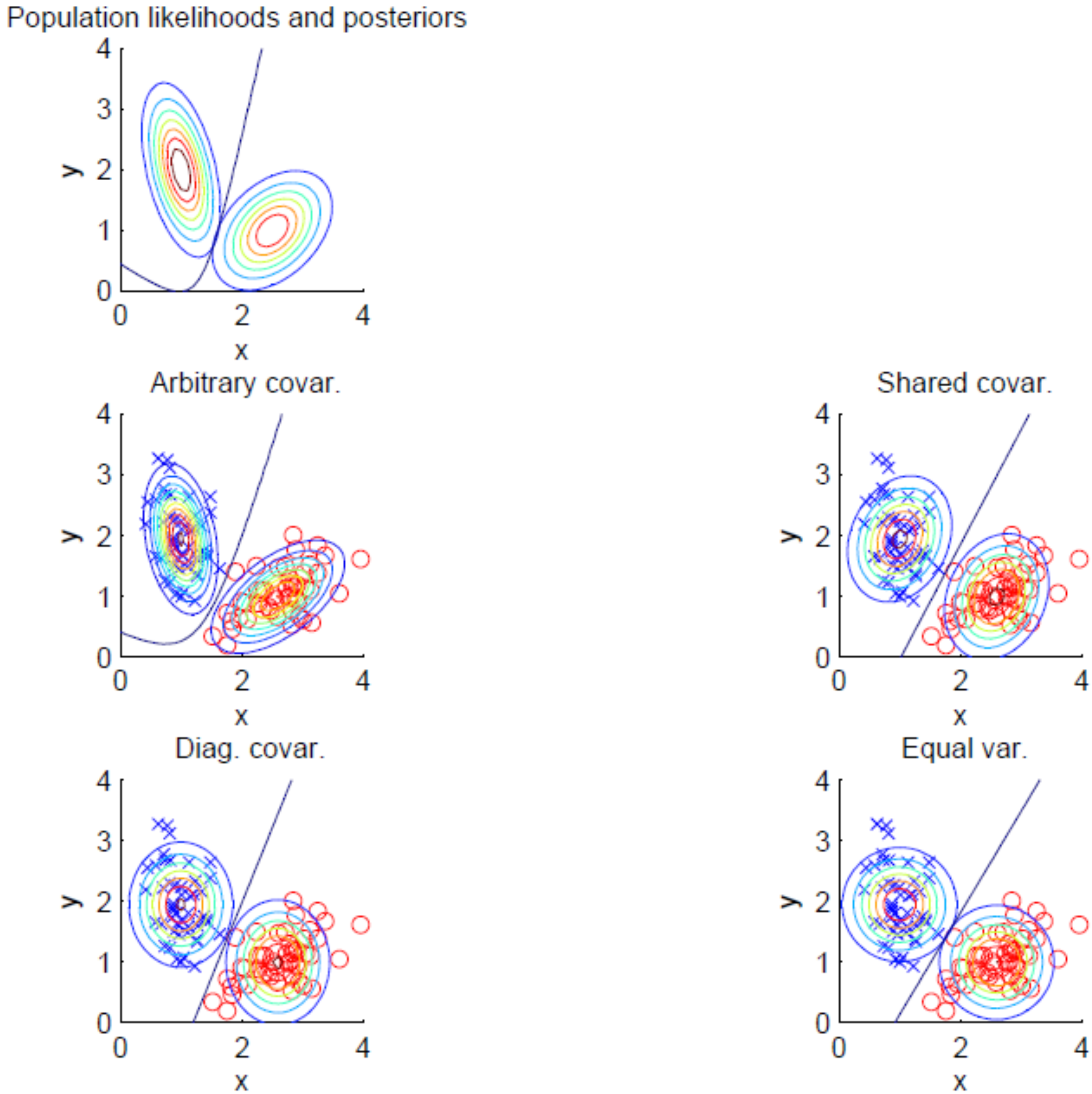
Model Selection

Assumption	*Covariance matrix*	*No of parameters*
Shared, Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	d
Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$d(d+1)/2$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ d(d+1)/2$

- As we increase complexity (less restricted **S**), bias decreases and variance increases
- Assume simple models (allow some bias) to control variance (regularization)

Diagonal **S**, equal variances





E. Alpaydin, Introduction to Machine Learning

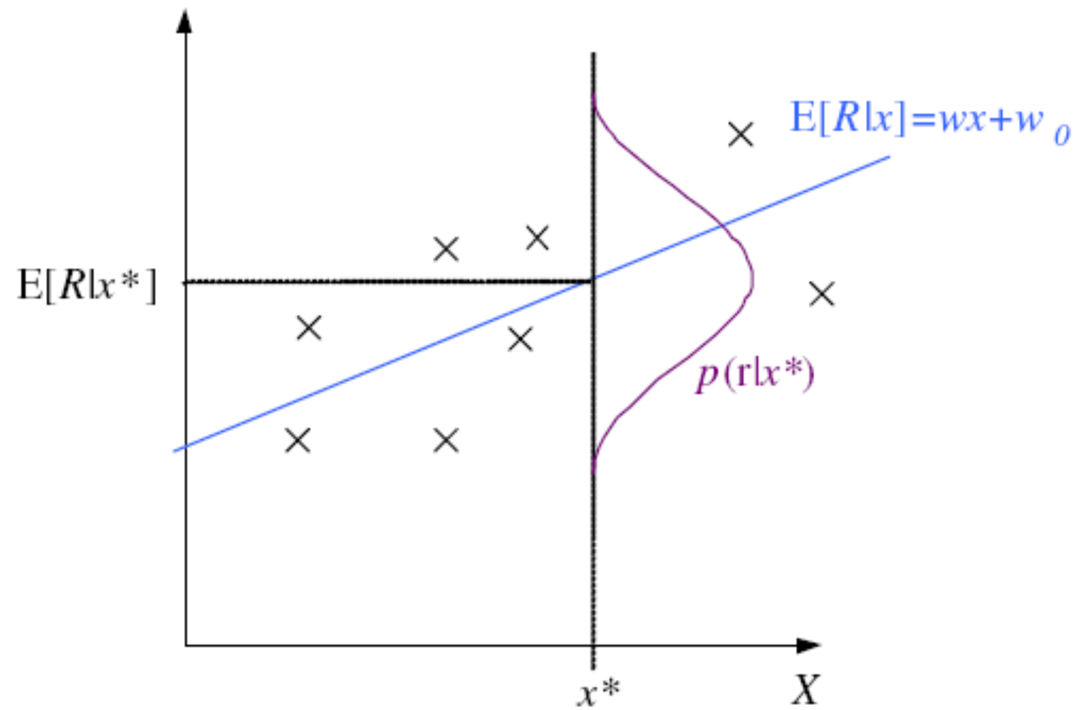
Regression

$$r = f(x) + \varepsilon$$

$$\text{estimator} : g(x \mid \theta)$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$p(r \mid x) \sim \mathcal{N}(g(x \mid \theta), \sigma^2)$$



$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} p(x^t, r^t)$$

$$= \log \prod_{t=1}^{N} p(r^t \mid x^t) + \log \prod_{t=1}^{N} p(x^t)$$

Regression: From LogL to Error

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{2\sigma^2} \right]$$

$$= -N \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

$$E(\theta \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

- Maximize the log likelihood is the same as minimize the error function.

Linear Regression $g\left(x^t \mid w_1, w_0\right) = w_1 x^t + w_0$

$$E\left(w_1, w_0 \mid \mathcal{X}\right) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - \left(w_1 x^t + w_0\right)\right]^2$$

$$\frac{\partial E\left(w_1, w_0 \mid \mathcal{X}\right)}{\partial w_0} = \sum_{t=1}^{N}\left[(r^t - w_1 x^t - w_0)(-1)\right] = 0 \Leftrightarrow \sum_t r^t = N w_0 + w_1 \sum_t x^t$$

$$\frac{\partial E\left(w_1, w_0 \mid \mathcal{X}\right)}{\partial w_1} = \sum_{t=1}^{N}\left[(r^t - w_1 x^t - w_0)(-x^t)\right] = 0 \Leftrightarrow \sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2$$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix} \qquad \mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$$

Multivariate Regression

$$r^t = g\left(x^t \mid w_0, w_1, \ldots, w_d\right) + \varepsilon$$

- Multivariate linear model

$$w_0 + w_1 x_1^t + w_2 x_2^t + \cdots + w_d x_d^t$$

$$E\left(w_0, w_1, \ldots, w_d \mid \mathcal{X}\right) = \frac{1}{2}\sum_t \left[r^t - w_0 - w_1 x_1^t - \cdots - w_d x_d^t\right]^2$$

Multivariate Regression

$$E\left(w_0, w_1, \ldots, w_d \mid \mathcal{X}\right) = \frac{1}{2}\sum_t \left[r^t - w_0 - w_1 x_1^t - \cdots - w_d x_d^t\right]^2$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}, X = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & & \\ 1 & x_1^N & x_2^N & \cdots & x_d^N \end{bmatrix}$$

$$\min_w \|\mathbf{r} - Xw\|^2$$

$$\frac{\partial \|\mathbf{r} - Xw\|^2}{\partial w} = -2X^T(\mathbf{r} - Xw) = 0 \Rightarrow w = (X^T X)^{-1} X^T \mathbf{r}$$

Multivariate Regression

$$E\left(w_0, w_1, \ldots, w_d \mid \mathcal{X}\right) = \frac{1}{2}\sum_t \left[r^t - w_0 - w_1 x_1^t - \cdots - w_d x_d^t\right]^2$$

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}, X = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & & \\ 1 & x_1^N & x_2^N & \cdots & x_d^N \end{bmatrix}$$

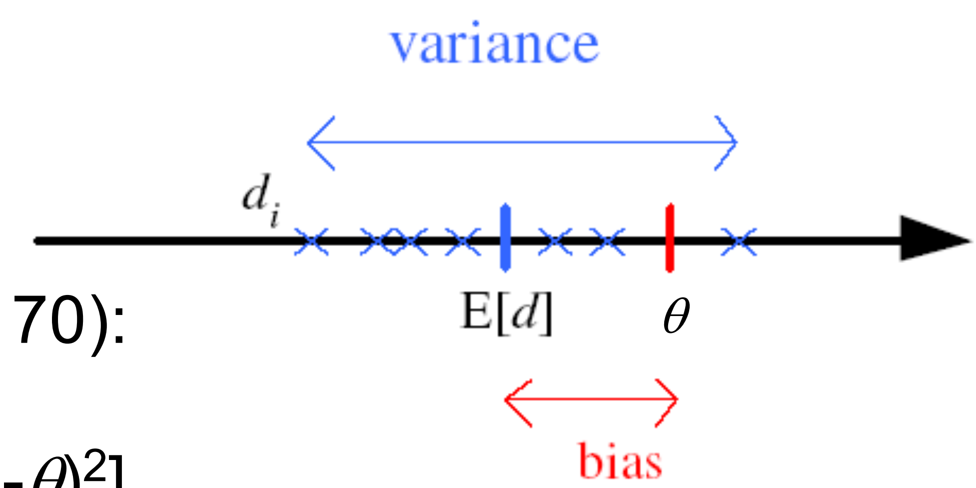
$$\min_w \|\mathbf{r} - Xw\|^2$$

$$\frac{\partial \|\mathbf{r} - Xw\|^2}{\partial w} = -2X^T(\mathbf{r} - Xw) = 0 \Rightarrow w = (X^T X)^{-1} X^T \mathbf{r}$$

Evaluating an Estimateor: Bias and Variance

Unknown parameter θ, Estimator $d_i = d(X_i)$ on sample X_i

Bias: $b_\theta(d) = E[d] - \theta$
Variance: $E[(d - E[d])^2]$



Mean square error (page 70):

$$
\begin{aligned}
r(d,\theta) &= E[(d-\theta)^2] \\
&= E[(d - E[d] + E[d] - \theta)^2] \\
&= (E[d] - \theta)^2 + E[(d - E[d])^2] \\
&= \text{Bias}^2 + \text{Variance}
\end{aligned}
$$

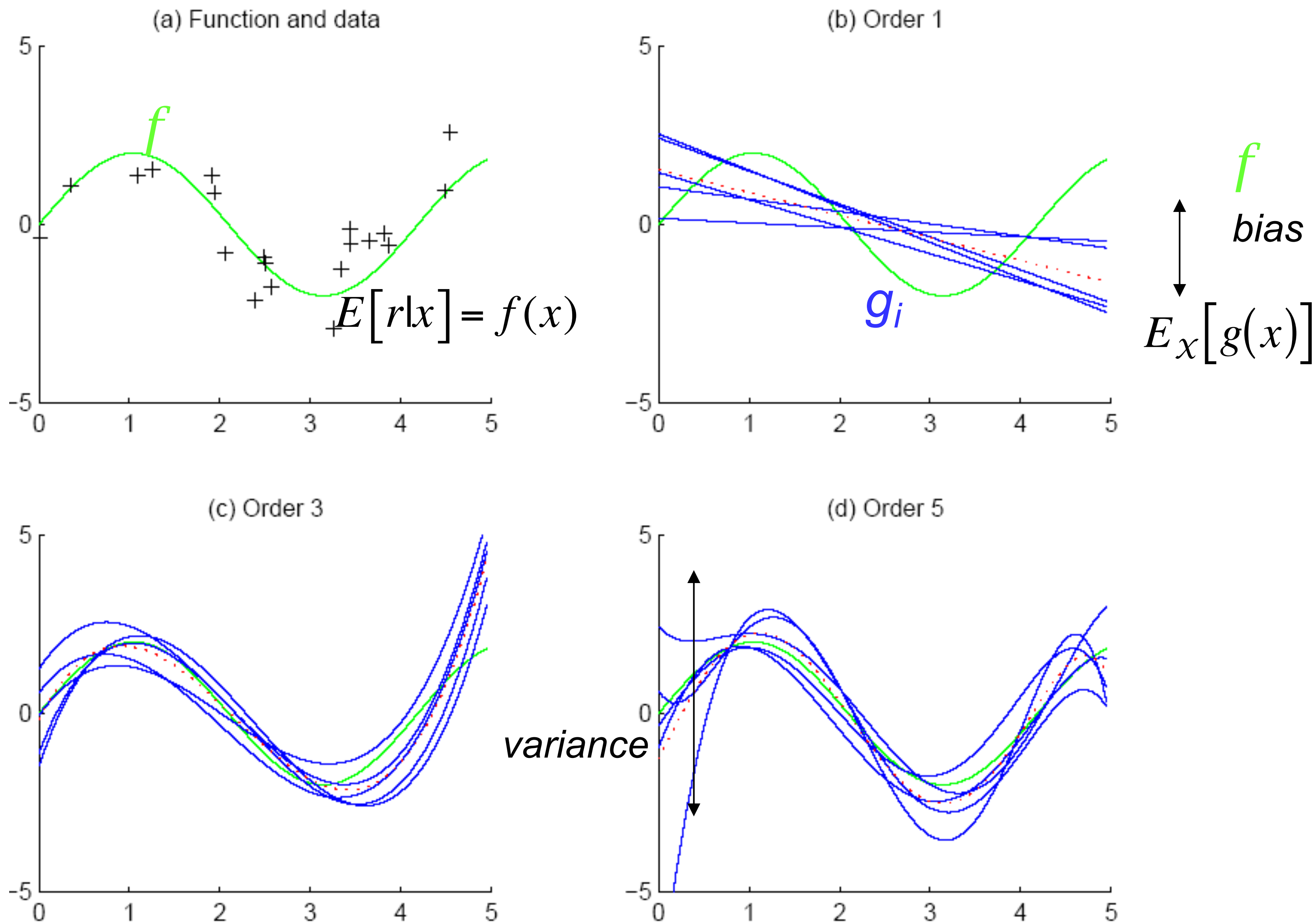
Bias/Variance Dilemma

- Example: $g_i(x)=2$

  has no variance and high bias

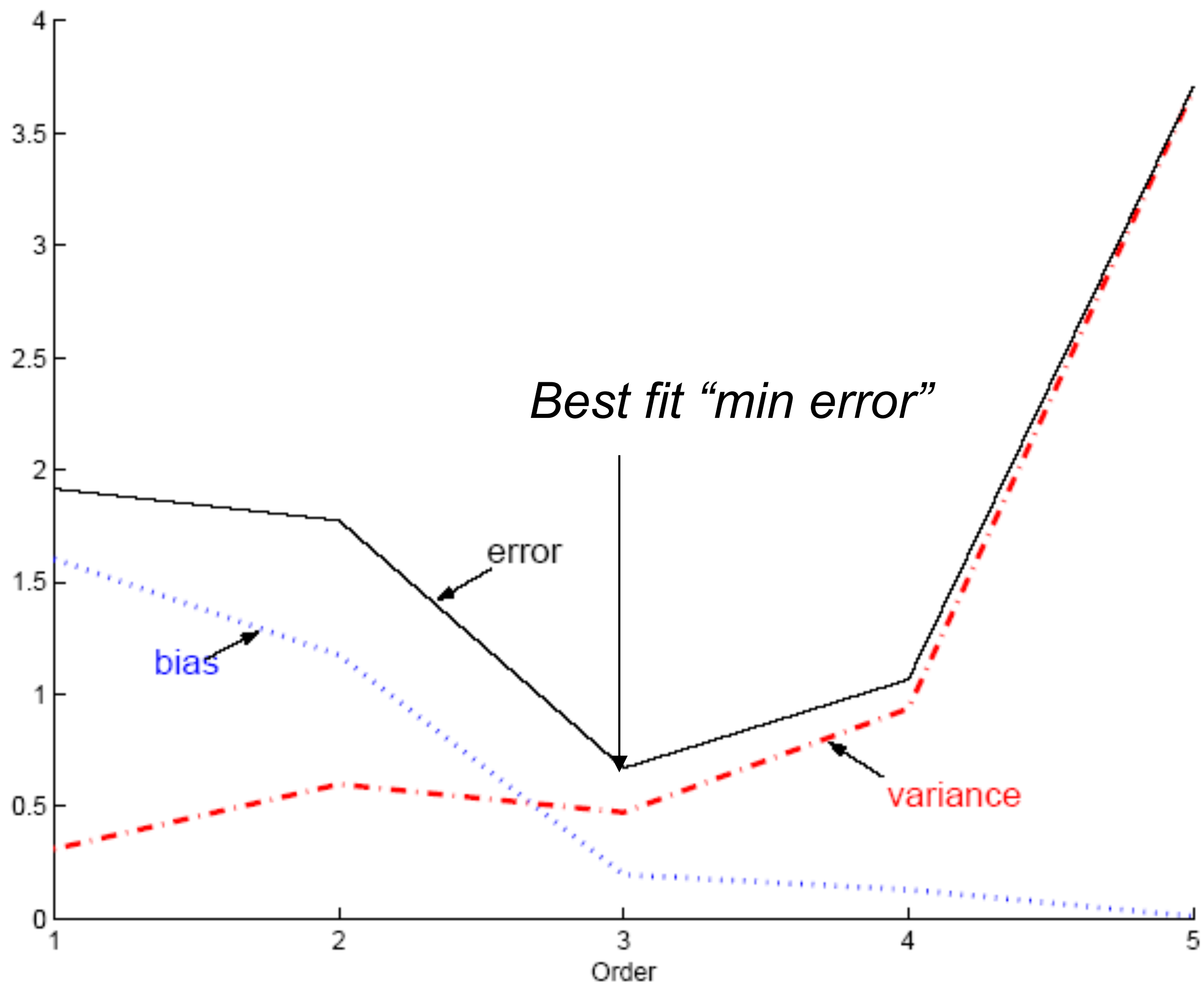
  $g_i(x)= \sum_t r^t_i/N$

  has lower bias with variance
- As we increase complexity, bias decreases (a better fit to data) and variance increases (fit varies more with data)
- Bias/Variance dilemma

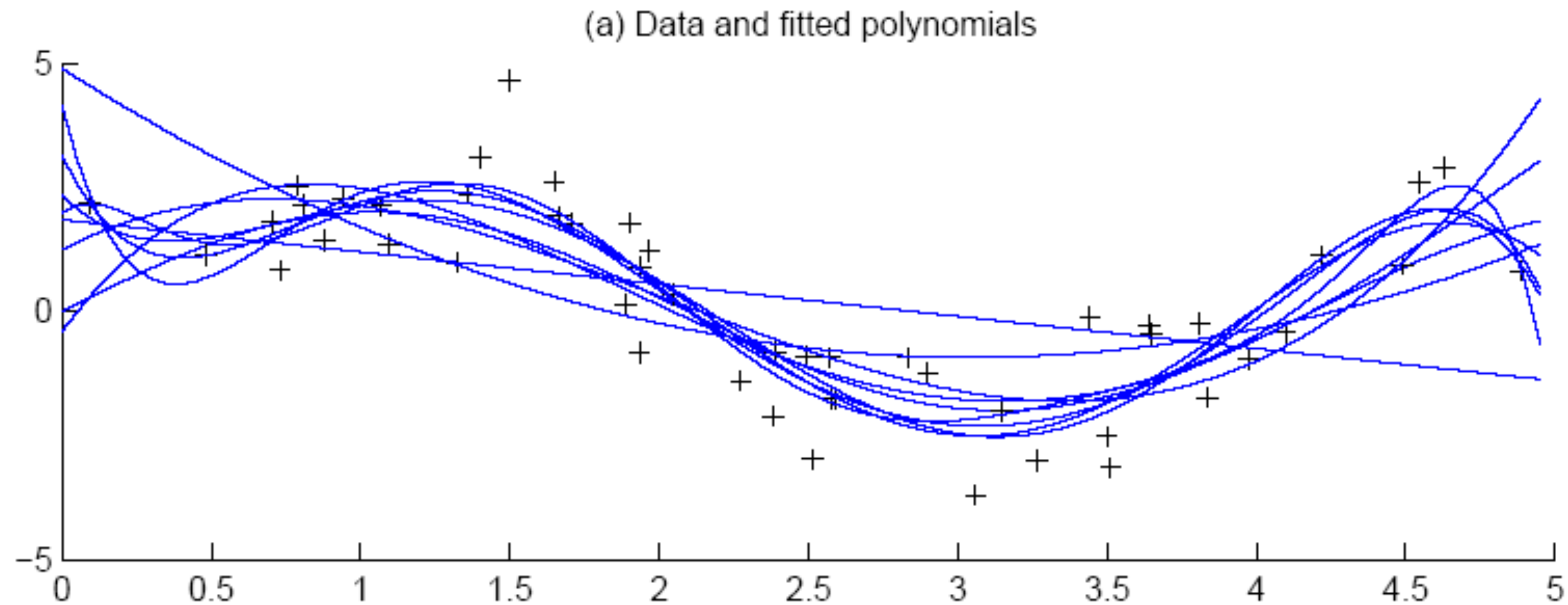


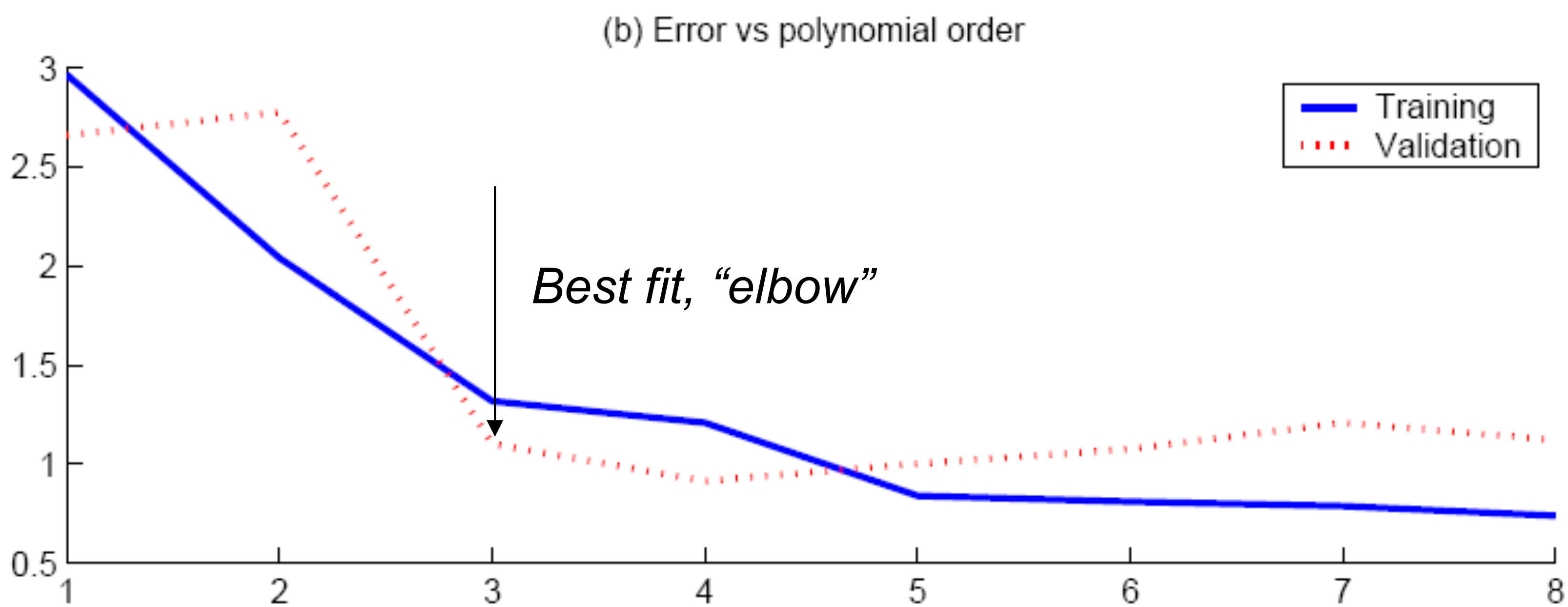
(a) Function and data
(b) Order 1
(c) Order 3
(d) Order 5
f
$E[r|x] = f(x)$
g_i
bias
$E_X[g(x)]$
variance

Polynomial Regression



E. Alpaydin, Introduction to Machine Learning





E. Alpaydin, Introduction to Machine Learning

Model Selection

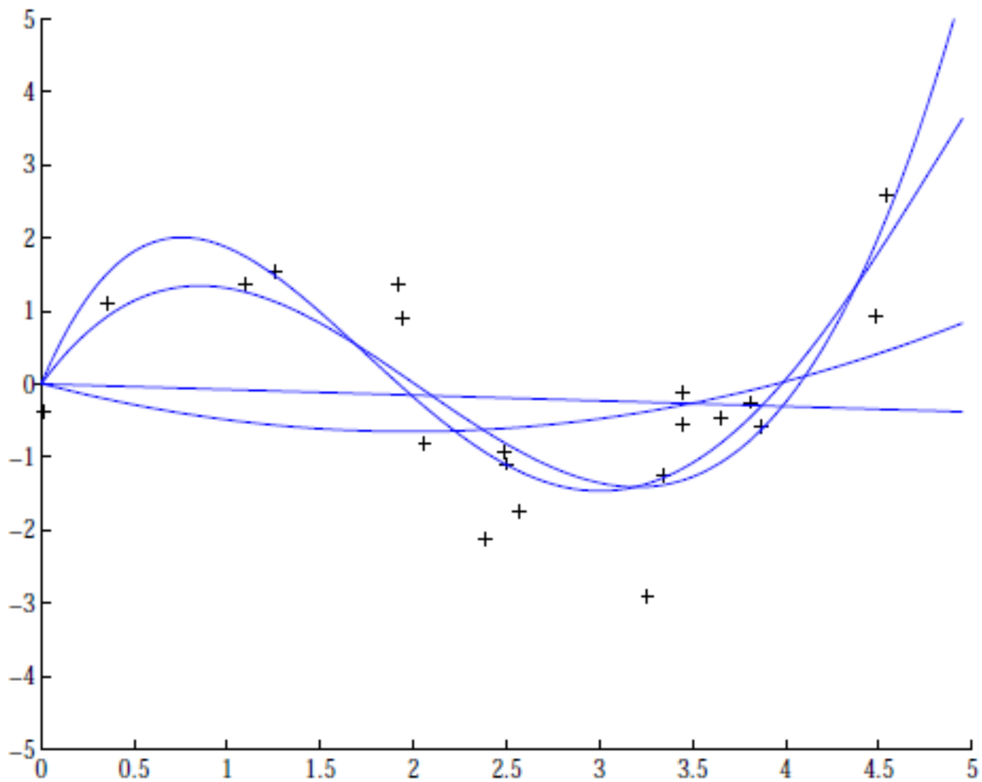
- Cross-validation: Measure generalization accuracy by testing on data unused during training
- Regularization: Penalize complex models

  E'=error on data + λ model complexity

- Regression with penalty on w:

$$E(\theta \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid w\right)\right]^2 + \lambda \sum_i w_i^2$$

Regression example



Coefficients increase in magnitude as order increases:

1: [-0.0769, 0.0016]

2: [0.1682, -0.6657, 0.0080]

3: [0.4238, -2.5778, 3.4675, -0.0002

4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]

$$regularization : E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \mathbf{w}\right)\right]^2 + \lambda \sum\nolimits_i w_i^2$$

E. Alpaydin, Introduction to Machine Learning

Regression example

$$\text{Regularization}: E(\mathbf{w}|\mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t|\mathbf{w}\right)\right]^2 + \lambda\sum_i w_i^2$$

When λ is small ($\rightarrow 0$), the model is unregularized and achieve 0 training error with complex model.

When λ is larger ($+\infty$), $g(x) = 0$ large training error with simplest model.