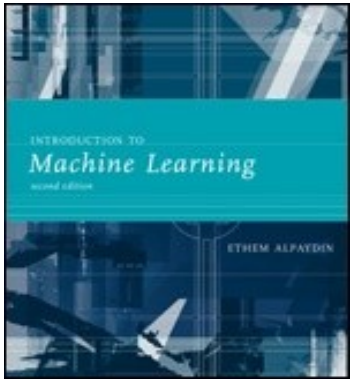
Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition



ETHEM ALPAYDIN, modified by Leonardo Bobadilla and some parts from http://www.cs.tau.ac.il/~apartzin/MachineLearning/

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2m

Outline

Last Class: Ch 4: Parametric Methods

Maximum Likelihood Estimation

Evaluating an estimator: Bias and Variance

This class: Ch 4: Parametric Methods

- The Bayes Estimator
- Parametric Classification
- Regression
-

CHAPTER 4:

Parametric Methods

Bias and Variance

Mean square error:

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

Variance: $E[(d-E[d])^2]$

Optimal Estimators

- Estimators are random variables as depend on random sample from the distribution
- We look for estimators (functions of samples) that have minimum expected square error
- Expected square error can be decomposed into bias (drift) and variance
- Sometimes accept larger errors but require unbiased estimators.

Bayes estimator

- Sometimes have some prior information on the possible value range that parameter may take
- Can’t have exact value but can provide a probability for each value (density function)
- Probability of a parameter before looking at data

Bayes Estimator

we are told that θ is approximately normal and with 90 percent confidence, θ lies between 5 and 9, symmetrically around 7. Then we can write $p(\theta)$ to be normal with mean 7 and because

$$P\{-1.64 < \frac{\theta - \mu}{\sigma} < 1.64\} = 0.9$$

$$P\{\mu - 1.64\sigma < \theta < \mu + 1.64\sigma\} = 0.9$$

we take $1.64\sigma = 2$ and use $\sigma = 2/1.64$. We can thus assume $p(\theta) \sim \mathcal{N}(7, (2/1.64)^2)$.

Bayes Estimator

$$p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)p(\theta)}{p(\mathcal{X})} = \frac{p(\mathcal{X}|\theta)p(\theta)}{\int p(\mathcal{X}|\theta')p(\theta')d\theta'}$$

- Assume this function have a peak around true value of the parameter

- Maximum a posteriori estimate will minimize error

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{X})$$

Bayes' Estimator

- Treat θ as a random var with prior $p(\theta)$
- Bayes' rule: $p(\theta|X) = p(X|\theta)\, p(\theta) / p(X)$
- Full: $p(x|X) = \int p(x|\theta)\, p(\theta|X)\, d\theta$
- Maximum a Posteriori (MAP): $\theta_{MAP} = \text{argmax}_\theta\, p(\theta|X)$
- Maximum Likelihood (ML): $\theta_{ML} = \text{argmax}_\theta\, p(X|\theta)$
- Bayes': $\theta_{Bayes'} = E[\theta|X] = \int \theta\, p(\theta|X)\, d\theta$

Bayes' Estimator: Example

- $x^t \sim \mathrm{N}(\theta, \sigma_o^2)$ and $\theta \sim \mathrm{N}(\mu, \sigma^2)$
- $\theta_{\mathrm{ML}} = m$
- $\theta_{\mathrm{MAP}} =$

$$\frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$$

Parametric Classification

- Bayes $P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)} = \frac{p(x|C_i)P(C_i)}{\sum_{k=1}^{K} p(x|C_k)P(C_k)}$

and use the discriminant function

$$g_i(x) = p(x|C_i)P(C_i)$$

or equivalently

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

Assumptions

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

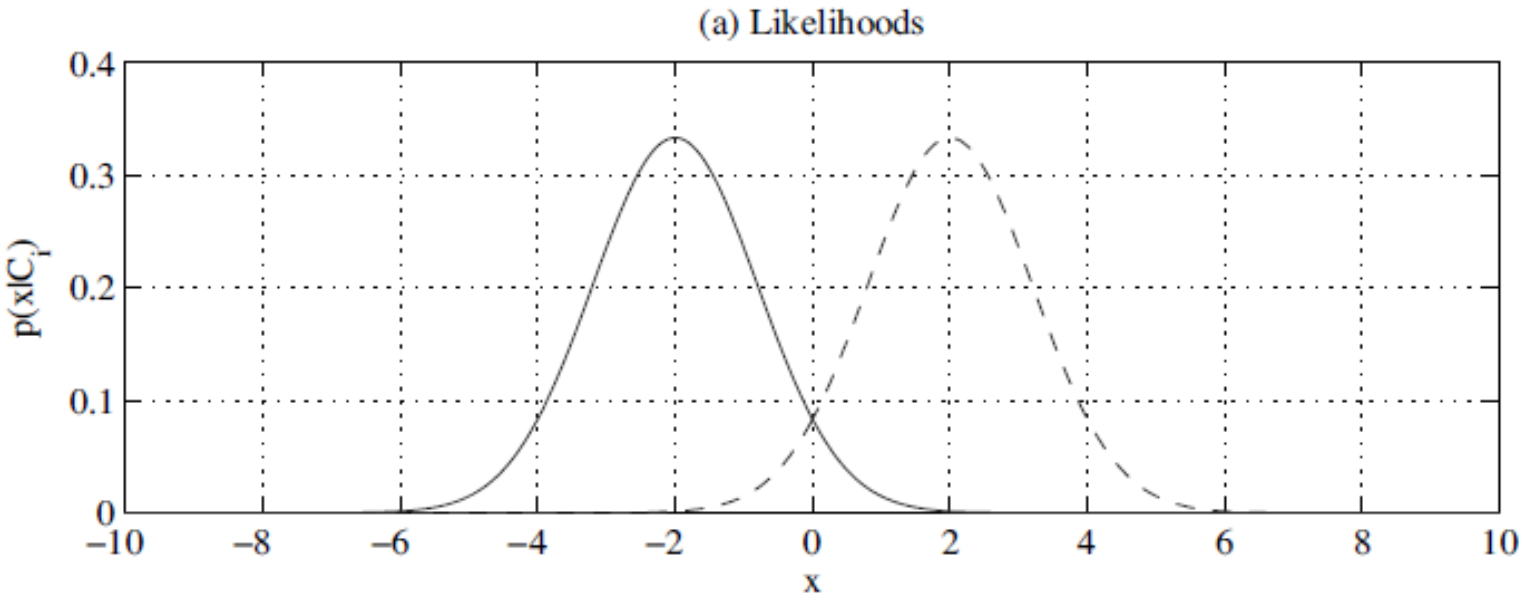
$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

Example

- Input: customer income
- Output: which car out of K models he will prefer
- Assume that for each model there is a group of customers with certain income
- Income in a single class distributed normal

$$\mathcal{X} = \{\boldsymbol{x}^t, \boldsymbol{r}^t\}_{t=1}^N \qquad r_i^t = \begin{cases} 1 & \text{if } \boldsymbol{x}^t \in C_i \\ 0 & \text{if } \boldsymbol{x}^t \in C_k, k \neq i \end{cases}$$

Example: True distributions



Example: Estimate parameters

$$m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t}$$

$$s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$$

Example: Discriminant

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

- Evaluate on a new input
- Select class with largest discriminant
- Assume: priors and variances are equal:

$$g_i(x) = -(x - m_i)^2$$

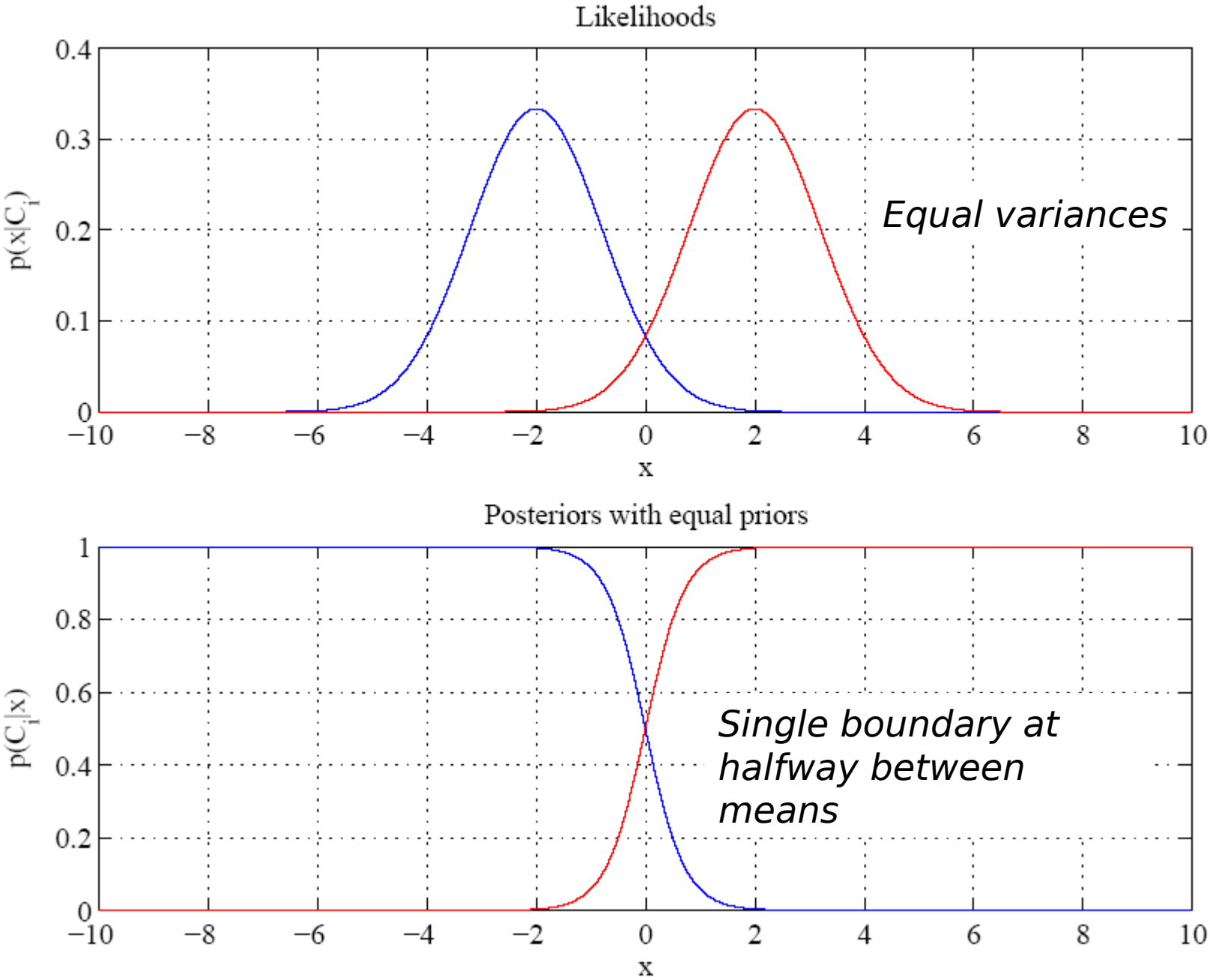
$$\text{Choose } C_i \text{ if } |x - m_i| = \min_k |x - m_k|$$

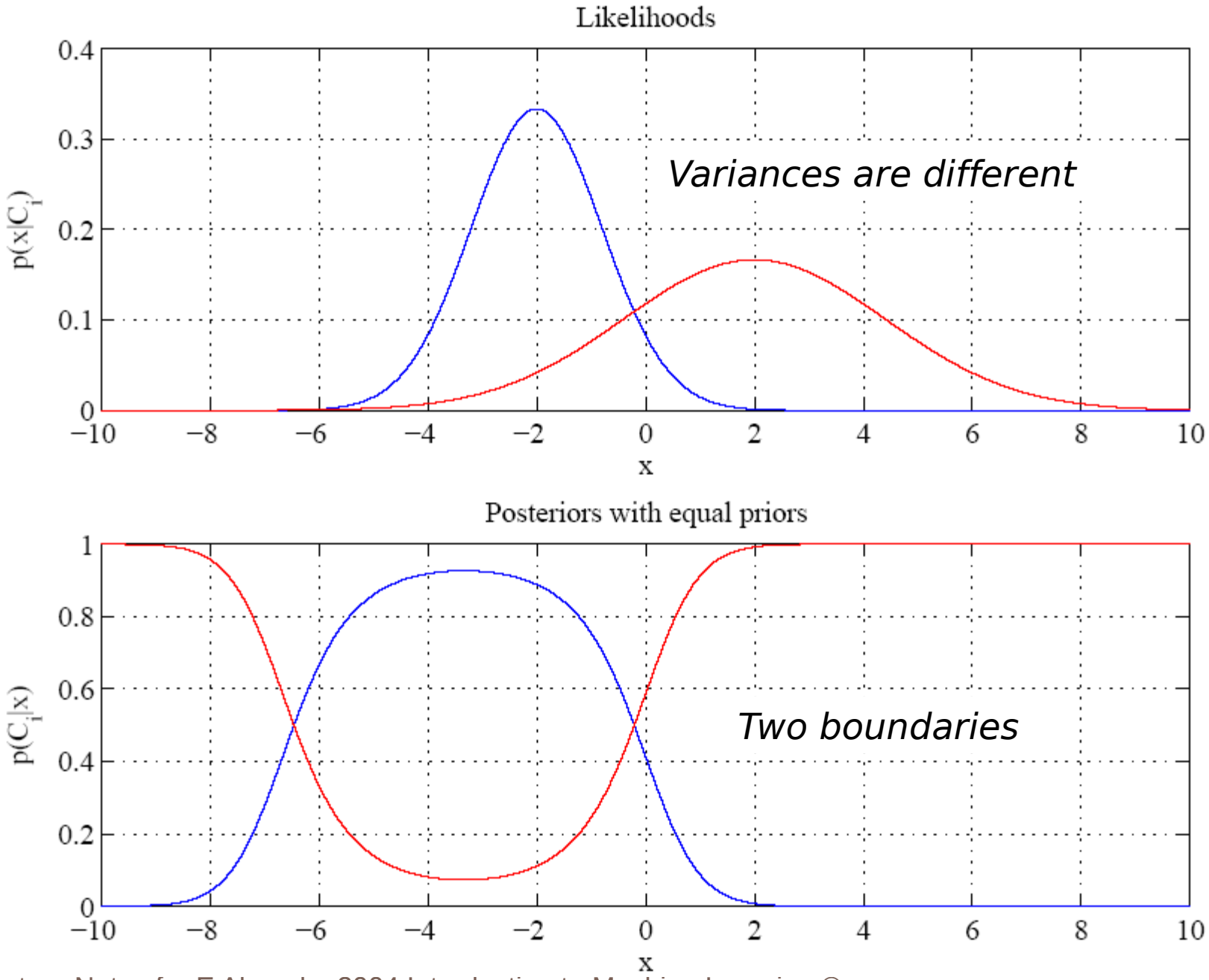
Boundary between classes

$$g_1(x) = g_2(x)$$

$$(x - m_1)^2 = (x - m_2)^2$$

$$x = \frac{m_1 + m_2}{2}$$


Likelihoods
p(x|C_i)
x
Equal variances
Posteriors with equal priors
p(C_i|x)
x
Single boundary at halfway between means


Likelihoods
p(x|C_i)
x
Variances are different
Posteriors with equal priors
p(C_i|x)
x
Two boundaries

Two approaches

- Likelihood-based approach (till now)
  - Calculate probabilities using Bayes rule
  - Compute discriminant function

- Discriminant function approach(later)
  - Directly estimate discriminant
  - Bypass probabilities estimation
  - Boundary between classes?

Regression

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

- x is independent variable, r is dependant variable
- Unknown f, want to approximate to predict future values
- Parametric approach: assume model with small number of parameters $g(x|\theta)$
- Find best parameters from data
- Also have to make assumption on noise

Regressions

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \qquad r = f(x) + \epsilon$$

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

- Have a training data (x,r)
- Find parameters to maximize likelihood
- In other words, what parameters makes data most probable

Regressions

$$p(x, r) = p(r|x)p(x)$$

$$\begin{aligned}\mathcal{L}(\theta|\mathcal{X}) &= \log \prod_{t=1}^{N} p(x^t, r^t) \\ &= \log \prod_{t=1}^{N} p(r^t|x^t) + \log \prod_{t=1}^{N} p(x^t)\end{aligned}$$

- Ignore the last term,(does not depend on parameters

Regression

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

- Minimize last term

Least Square Estimate

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

- Minimize this