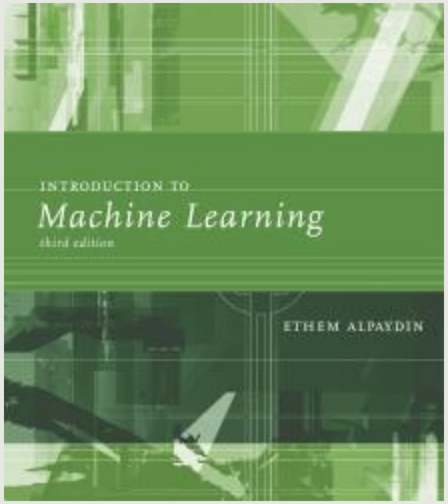


Lecture Slides for

INTRODUCTION TO MACHINE LEARNING 3RD EDITION

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml3e

CHAPTER 3: BAYESIAN DECISION THEORY

Probability and Inference

- Result of tossing a coin is $\in$ {Heads, Tails}
- Random var $X \in \{1,0\}$

  Bernoulli: $P\{X=1\} = p_o^{X} (1 - p_o)^{(1-X)}$
- Sample: $\mathbf{X} = \{x^t\}_{t=1}^{N}$

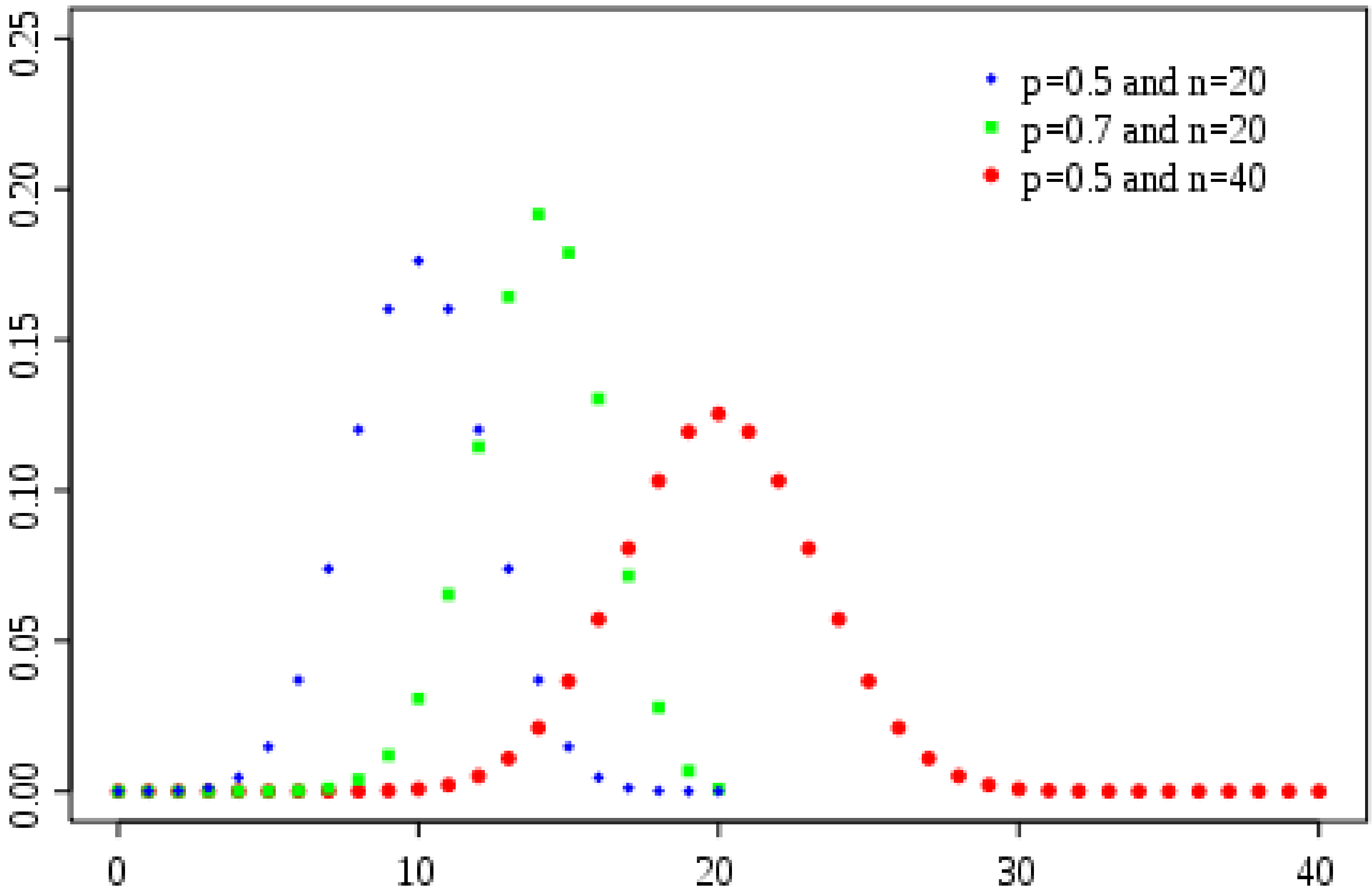
  Estimation: $p_o = \#\{\text{Heads}\}/\#\{\text{Tosses}\} = \sum_t x^t / N$
- Prediction of next toss:

  Heads if $p_o > ½$, Tails otherwise

In the theory of probability and statistics, a **Bernoulli trial** is an experiment whose outcome is random and can be either of two possible outcomes, "success" and "failure".

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Binomial Distribution



Classification

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

$$\text{Choose } \begin{cases} C = 1 \text{ if } P(C = 1 | x_1, x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

or

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

Bayes' Rule

prior *likelihood*

posterior

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

evidence

$$P(\mathrm{C}=0) + P(\mathrm{C}=1) = 1$$

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

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

Bayes' Rule: K>2 Classes

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

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

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

Choose C_i if $P(C_i|\mathbf{x}) = \max_k P(C_k|\mathbf{x})$

Remember: The disease/symptom example

Losses and Risks

- Actions: α_i
- Loss of α_i when the state is C_k : λ_{ik}
- Expected risk (Duda and Hart)

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x})$$

$$\text{Choose } \alpha_i \text{ if } R(\alpha_i|\mathbf{x}) = \min_k R(\alpha_k|\mathbf{x})$$

Remark:

λ_{ik} is the cost of choosing *i* when *k* is correct!

If we use accuracy/error, then

λ_{ik} := If i=k then 0 else 1!

Losses and Risks: 0/1 Loss

$$\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$$

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{k=1}^{K} \lambda_{ik} P(C_k|\mathbf{x}) \\ &= \sum_{k \neq i} P(C_k|\mathbf{x}) \\ &= 1 - P(C_i|\mathbf{x}) \end{aligned}$$

For minimum risk, choose the most probable class

Losses and Risks: Reject

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

$$R\left(\alpha_{K+1}|\mathbf{x}\right) = \sum_{k=1}^{K} \lambda P\left(C_k|\mathbf{x}\right) = \lambda$$

$$R\left(\alpha_i|\mathbf{x}\right) = \sum_{k \neq i} P\left(C_k|\mathbf{x}\right) = 1 - P\left(C_i|\mathbf{x}\right)$$

The Optimum Decision Rule

Choose C_i if $R(\alpha_i|\mathbf{x}) < R(\alpha_k|\mathbf{x})$ for all $k \neq i$ and

$R(\alpha_i|\mathbf{x}) < R(\alpha_{K+1}|\mathbf{x})$

Reject if $R(\alpha_{K+1}|\mathbf{x}) < R(\alpha_i|\mathbf{x}), i = 1, \ldots, K$

Given the loss function $\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$

Choose C_i if $P(C_i|\mathbf{x}) > P(C_k|\mathbf{x}) \ \forall k \neq i$ and $P(C_i|\mathbf{x}) > 1 - \lambda$

Reject otherwise

Example:

C_1=has cancer
C_2=has not cancer
$\lambda_{12}=9$
$\lambda_{21}=72$
Homework:
a) Determine the optimal decision making strategy
Inputs: $P(C_1|x)$, $P(C_2|x)$
Decision Making Strategy:…
b) Now assume we also have a reject option and the cost for making no decision are 3:
$\lambda_{reject,\,2}=3$
$\lambda_{reject,\,1}=3$
Inputs: $P(C_1|x)$, $P(C_2|x)$
Decision Making Strategy: …

a) Determine the optimal decision making strategy
Inputs: $P(C_1|x)$, $P(C_2|x)$

$R(a_1|x)=9\times P(C_2|x)$; $R(a_2|x)=72\times P(C_1|x)$
$R(a_{reject}|x)=3$
Setting those equal receive:
$9\times P(C_2|x)=72\times P(C_1|x) \leftrightarrow (P(C_2|x)/P(C_1|x))=8$; additionally using $P(C_1|x)+P(C_2|x)=1$ we receive: $P(C_1|x)=1/9$ and $P(C_2|x)=8/9$ and the risk-minimizing decision rule becomes:
IF $P(C_1|x)>1/9$ THEN choose C_1 ELSE choose C_2
b) Now assume we also have a reject option and the cost for making no decision are 3:
$\lambda_{reject,2}=3$
$\lambda_{reject,\ 1}=3$
Input: $P(C_1|x)$
First we find equating $R(a_{reject}|x)$ with $R(a_1|x)$ and $R(a_2|x)$:
If $P(C_2|x)\geq 1/3 \leftrightarrow P(C_1|x) \leq 2/3$ reject should be preferred over class1 and $P(C_1|x)\geq 1/24$ reject should be preferred over class2. Combining this knowledge with the previous decision rule we receive:

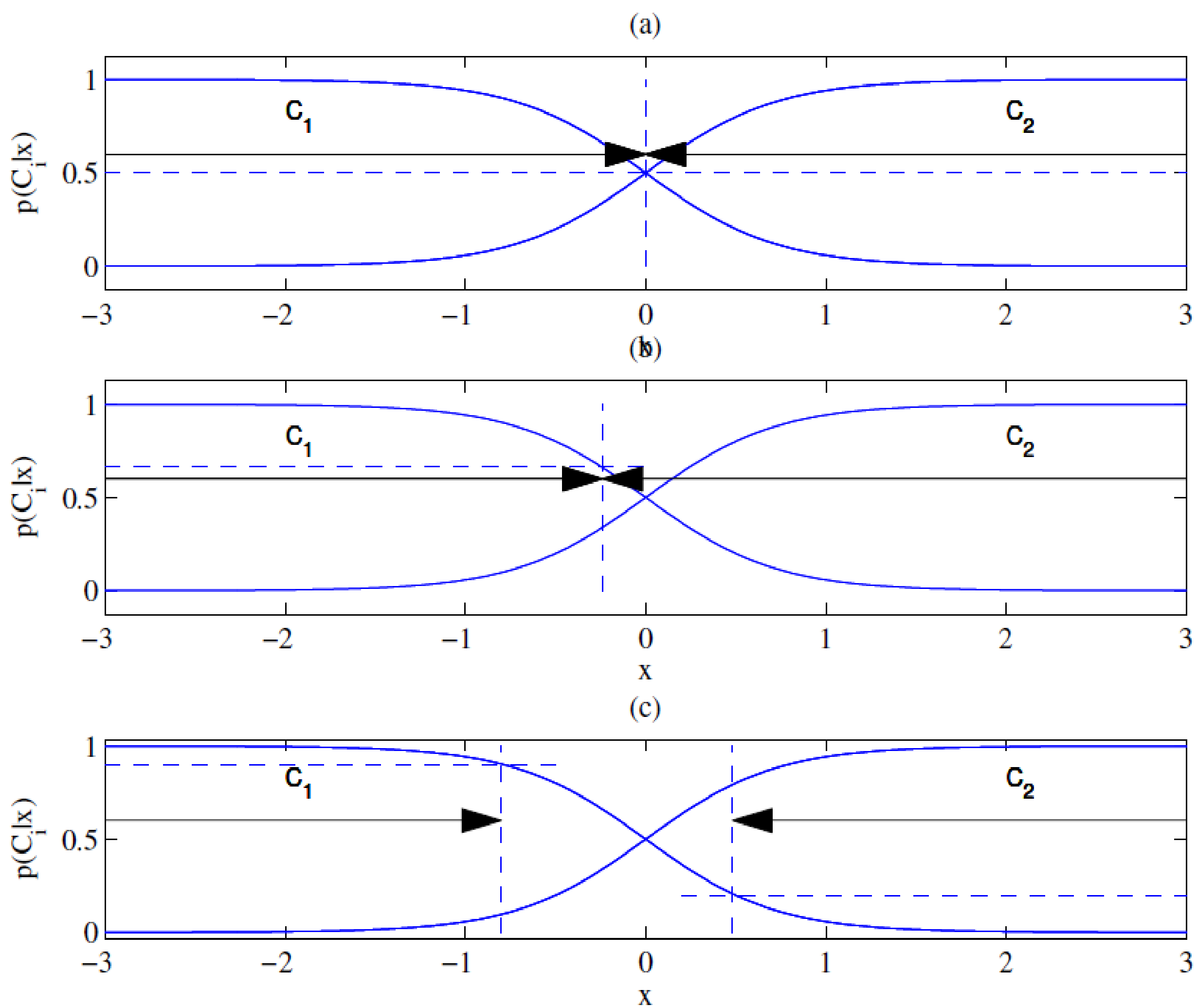
IF $P(C_1|x)\in[0,1/24]$ THEN choose class2
ELSE IF $P(C1|x)\in[2/3,1]$ THEN choose class1
ELSE choose reject

Different Losses and Reject

Equal losses

Unequal losses

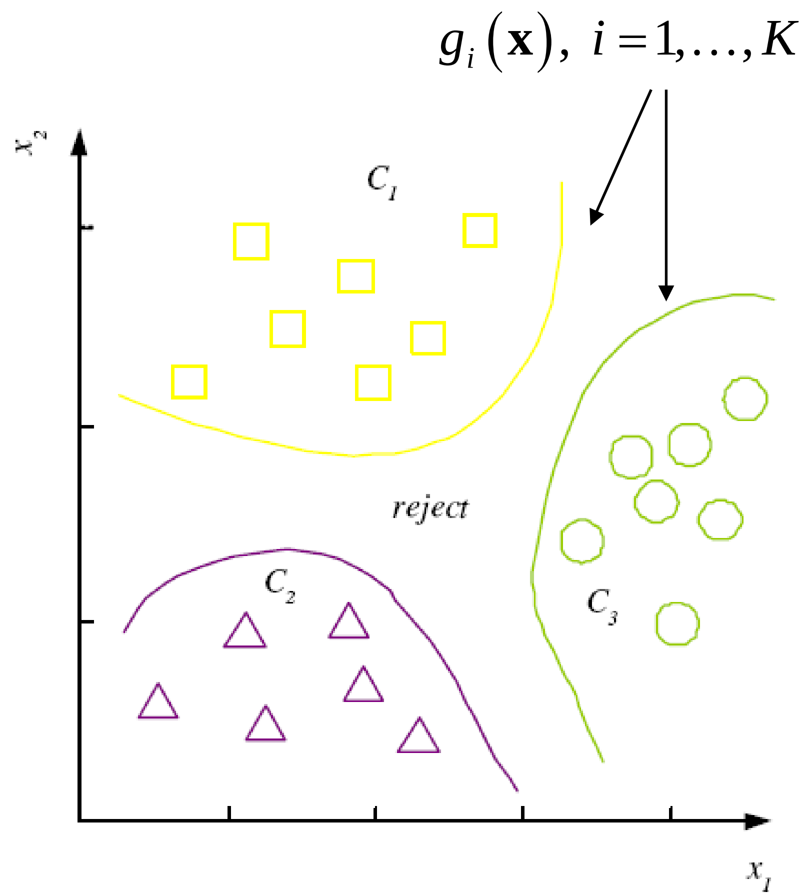
With reject



Discriminant Functions

choose C_i if $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i|\mathbf{x}) \\ P(C_i|\mathbf{x}) \\ p(\mathbf{x}|C_i)P(C_i) \end{cases}$$



K decision regions $R_1,\ldots,R_K$

$$R_i = \{\mathbf{x}|g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

K=2 Classes

- Dichotomizer (K=2) vs Polychotomizer ($K > 2$)
- $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

$$\text{choose} \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

- *Log odds:* $\log \dfrac{P(C_1|\mathbf{x})}{P(C_2|\mathbf{x})}$

Association Rules

- Association rule: An association rule is an implication of the form $X \rightarrow Y$
- *People who buy/click/visit/enjoy X are also likely to buy/click/visit/enjoy Y.*
- A rule implies association, not necessarily causation.

Association measures

- Support of $(X \longrightarrow Y)$:

$$P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- Confidence of $(X \longrightarrow Y)$:

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

- Lift or interest of $(X \longrightarrow Y)$:

$$Lift(X \to Y) \equiv \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)} = \frac{P(X \mid Y)}{P(X)}$$

Support shows the statistical significance of the rule, whereas confidence shows the strength of the rule.

If lift $> 1, \rightarrow$ X makes Y more likely,
If lift $< 1, \rightarrow$ X makes Y less likely.

Example:

Transaction	Items in basket
1	milk, bananas, chocolate
2	milk, chocolate
3	milk, bananas
4	chocolate
5	chocolate
6	milk, chocolate

SOLUTION:

$$\begin{aligned}
\text{milk} \rightarrow \text{bananas} &: \text{Support} = 2/6, \text{Confidence} = 2/4 \\
\text{bananas} \rightarrow \text{milk} &: \text{Support} = 2/6, \text{Confidence} = 2/2 \\
\text{milk} \rightarrow \text{chocolate} &: \text{Support} = 3/6, \text{Confidence} = 3/4 \\
\text{chocolate} \rightarrow \text{milk} &: \text{Support} = 3/6, \text{Confidence} = 3/5
\end{aligned}$$

Apriori algorithm (Agrawal et al., 1996)

- For (X,Y,Z), a 3-item set, to be frequent (have enough support), (X,Y), (X,Z), and (Y,Z) should be frequent.
- If (X,Y) is not frequent, none of its supersets can be frequent.
- Once we find the frequent *k*-item sets, we convert them to rules with enough confidence: X, Y→Z, ... and X → Y, Z, ...