

Lecture Slides for

INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

© The MIT Press, 2010

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml2e>

CHAPTER 3:

Bayesian Decision Theory

Probability and Inference

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

$$\text{Bernoulli: } P\{X=1\} = p_o^x (1 - p_o)^{(1-x)}$$

- Sample: $\mathbf{X} = \{x^t\}_{t=1}^N$

$$\text{Estimation: } p_o = \# \{\text{Heads}\} / \#\{\text{Tosses}\} = \sum_t x^t / N$$

- Prediction of next toss:

Heads if $p_o > 1/2$, Tails otherwise

Classification

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

$$\begin{array}{c} \textit{posterior} \\ \curvearrowright \\ P(C | \mathbf{x}) = \frac{\overset{\textit{prior}}{P(C)} \overset{\textit{likelihood}}{p(\mathbf{x} | C)}}{\underset{\textit{evidence}}{p(\mathbf{x})}} \end{array}$$

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

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

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

Bayes' Rule: $K > 2$ Classes

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

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

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

Losses and Risks

- Actions: α_i
- Loss of α_i when the state is C_k : λ_{ik}
- Expected risk (Duda and Hart, 1973)

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

choose α_i if $R(\alpha_i | \mathbf{x}) = \min_k R(\alpha_k | \mathbf{x})$

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

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

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

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

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

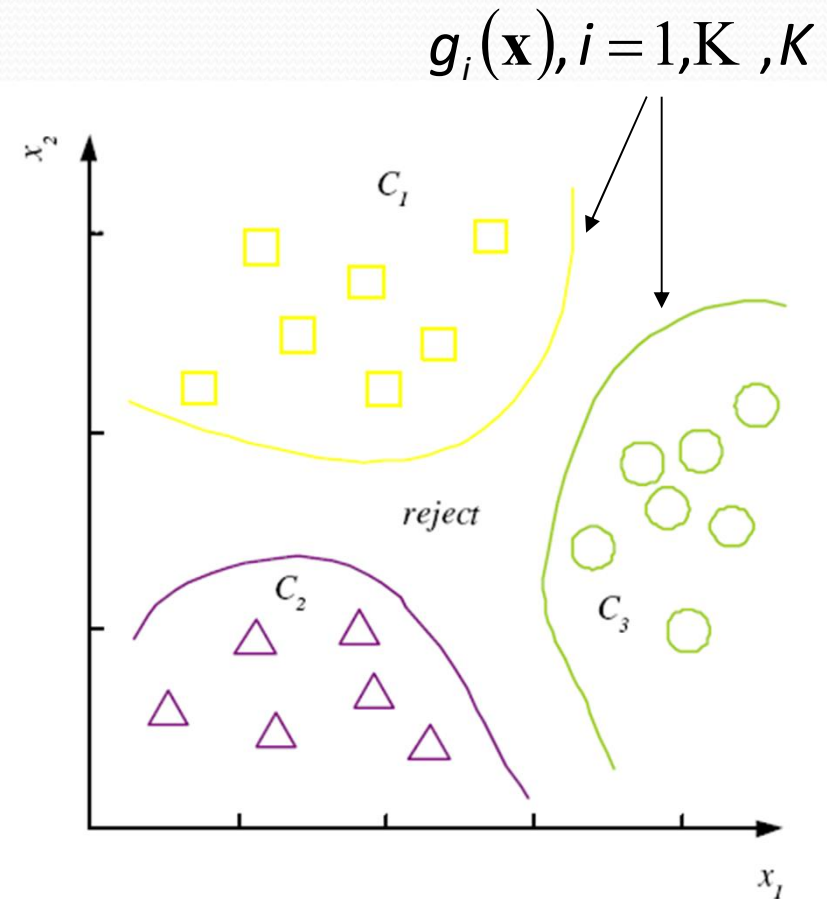
Discriminant Functions

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 $\mathcal{R}_1, \dots, \mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} | g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$



$K=2$ Classes

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

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

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

Utility Theory

- Prob of state k given evidence \mathbf{x} : $P(S_k | \mathbf{x})$
- Utility of α_i when state is k : U_{ik}
- Expected utility:

$$EU(\alpha_i | \mathbf{x}) = \sum_k U_{ik} P(S_k | \mathbf{x})$$

Choose α_i if $EU(\alpha_i | \mathbf{x}) = \max_j EU(\alpha_j | \mathbf{x})$

Association Rules

- Association rule: $X \rightarrow Y$
- *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 ($X \rightarrow Y$):

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

- Confidence ($X \rightarrow Y$):

$$P(Y | X) = \frac{P(X, Y)}{P(X)}$$

- Lift ($X \rightarrow Y$):
$$= \frac{\# \{ \text{customers who bought } X \text{ and } Y \}}{\# \{ \text{customers who bought } X \}}$$

$$= \frac{P(X, Y)}{P(X)P(Y)} = \frac{P(Y | X)}{P(Y)}$$

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: $X, Y \rightarrow Z, \dots$
and $X \rightarrow Y, Z, \dots$