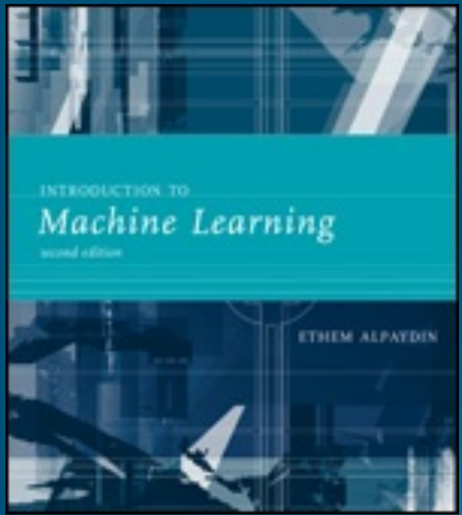
Lecture Slides for



INTRODUCTION TO

Machine Learning

2nd Edition

ETHEM ALPAYDIN

alpaydin@boun.edu.tr

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

In preparation of these slides, I have benefited from slides prepared by:

E. Alpaydin (Intro. to Machine Learning),

D. Bouchaffra and V. Murino (Pattern Classification and Scene Analysis),

R. Gutierrez-Osuna (Texas A&M)

A. Moore (CMU)

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: $\boldsymbol{X} = \{x^t\}_{t=1}^N$

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

  Heads if $p_o > ½$, Tails otherwise

Game

- You record the following tosses:
- {H, T, T, T, H, T, T, T, T, H, H, T, H, T, T, H, H, T, T, H?}
- You win if you get the next toss right.
- What do you guess?

- You win 10TL and lose 5TL if you guess the next toss right.
  - How do you compute your earnings?
- What do you guess?
  - Based on maximizing your earnings?

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 {0,1}
- Prediction:

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

or

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

Bayes' Rule

$$\underbrace{P(C \mid \mathbf{x})}_{\textit{posterior}} = \frac{\overbrace{P(C)}^{\textit{prior}}\,\overbrace{p(\mathbf{x} \mid C)}^{\textit{likelihood}}}{\underbrace{p(\mathbf{x})}_{\textit{evidence}}}$$

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

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

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

Game

P(x|hamsi)		
	short	tall
white	0.6	0.1
gray	0.2	0.1

P(x|lufer)		
	short	tall
white	0.05	0.2
gray	0.05	0.7

You caught a tall and white fish.

Is it hamsi or lufer?

Bayes' Rule: *K*>2 Classes

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

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

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

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

Losses and Risks

- Actions: α_i
- Loss of α_i when the state is C_k : λ_{ik}

	Predicted 0	Predicted 1
Actual 0	0	1
Actual1	1	0

e.g. cancer prediction

	Predicted 0	Predicted 1
Actual 0	0	1
Actual1	100	0

Losses and Risks

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

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

$$\text{choose } \alpha_i \text{ if } R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$$

Losses and Risks: 0/1 Loss

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

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

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

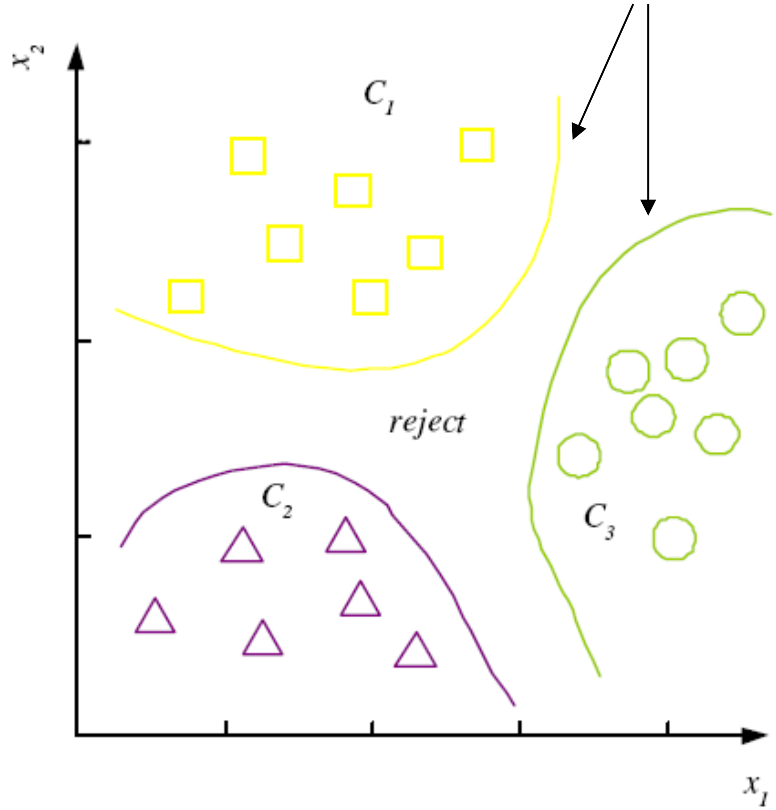
reject otherwise

Discriminant Functions

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

$$g_i(\mathbf{x}), i = 1, \ldots, K$$

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



K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$

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

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

Utility Theory

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

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

$$\text{Choose} \quad \alpha_i \text{ if } EU(\alpha_i \mid \mathbf{x}) = \max_j EU(\alpha_j \mid \mathbf{x})$$

- This is equivalent to minimizing the risk $R(\alpha_i|\mathbf{x})$
- Based on the specific problem, other functions might be optimized (e.g. Minimize worst possible loss, maximize money earned...)

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 ($X \rightarrow Y$):

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

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

- Lift ($X \rightarrow Y$):

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

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 → Z, ...

  and X → Y, Z, ...

See also the FP-Growth Algorithm:

Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In SIGMOD, 2000