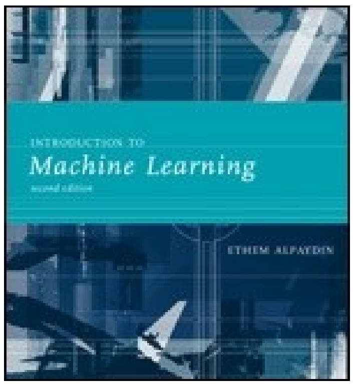
Lecture Slides for

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 2 Supervised Learning (Sec 2.1-2.4)

Learning Multiple Classes

Regression

Model Selection and Generalization

Dimensions of a Supervised Learning

This class:

- Bayes theorem
- Losses and risks
- Discriminant functions
- Association Rules

CHAPTER 3:

Bayesian Decision Theory

Making Decision Under Uncertainty

- Probability theory is the framework for making decisions under uncertainty.
- Use Bayes rule to calculate the probability of the classes
- Make rational decision among multiple actions to minimize expected risk
- Learning association rules from data

Unobservable variables

- Tossing a coin is completely random process, can’t predict the outcome
- Only can talk about the probabilities that the outcome of the next toss will be head or tails
- If we have access to extra knowledge (exact composition of the coin, initial position, force etc.) the exact outcome of the toss can be predicted

Unobservable Variable

- Unobservable variable is the extra knowledge that we don't have access to
- Coin toss: the only observable variables is the outcome of the toss
- $x=f(z)$, z is unobservables , x is observables
- f is deterministic function

Bernoulli Random Variable

- Result of tossing a coin is ∈ {Heads,Tails}
- Define a random variable $X \in \{1,0\}$
- p_o the probability of heads
- $P(X = 1) = p_o$ *and* $P(X = 0) = 1 - P(X = 1) = 1 - p_o$
- Assume asked to predict the next toss
- If know p_o we would predict heads if $p_o > 1/2$
- Choose more probable case to minimize probability of the error $1 - p_o$

Estimation

- What if we don't know P(X)
- Want to estimate from given data (sample) $\mathcal{X}$
- Realm of statistics
- Sample $\mathcal{X}$ generated from probability distribution of the observables x^t
- Want to build an approximator *p(x) using sample* $\mathcal{X}$
- In coin toss example: sample is outcomes of past N tosses and in distribution is characterized by single parameter p_o

Parameter Estimation

$$\hat{p}_o = \frac{\#\{\text{tosses with outcome heads}\}}{\#\{\text{tosses}\}}$$

$$\mathcal{X} = \{1, 1, 1, 0, 1, 0, 0, 1, 1\}$$

$$\hat{p}_o = \frac{\sum_{t=1}^{N} x^t}{N} = \frac{6}{9}$$

Classification

- Credit scoring: two classes – high risk and low risk
- Decide on observable information: (income and saving)
- Have reasons to believe that these 2 variable gives us idea about the credibility of a customer
- Represent by two random variable X_1 and X_2
- Can't observe customer intentions and moral codes
- Can observe credibility of a past customer
- Bernoulli random variable C conditioned on $X=[X_1 , X_2]^T$
- Assume we know $P(C| X_1 , X_2)$

Classification

- Assume know $P(C| X_1 , X_2)$
- New applications $X_1=x_1$, $X_2 =x_2$

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

or equivalently

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

Classification

`

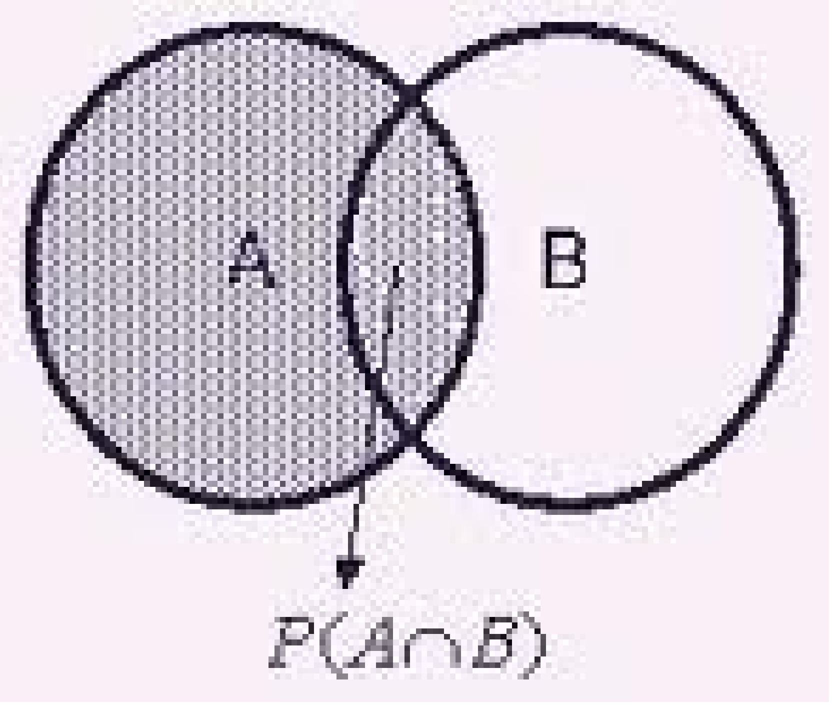
The probability of error is $1 - \max(P(C = 1|x_1, x_2), P(C = 0|x_1, x_2))$.

- Similar to coin toss but C is conditioned on two other observable variables $x=[x_1 ,x_2]^T$
- The problem : Calculate $P(C|x)$
- Use Bayes rule

Conditional Probability

- Probability of A (point will be inside A) if we know that B happens (point is inside B)

- $P(A|B)=P(A \cap B)/P(B)$



Bayes Rule

- $P(A|B)=P(A \cap B)/P(B)$
- $P(B|A)= P(A \cap B)/P(A) \Rightarrow P(A \cap B)=P(B|A)*P(A)$

  - **$P(A|B)=P(B|A)*P(A)/P(B)$**

Bayes Rule

prior *likelihood*

posterior

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

evidence

- **Prior**: probability of a customer is high risk regardless of x.
- Knowledge we have as to the value of C before looking at observables x

Bayes Rule

prior *likelihood*

posterior

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

evidence

- **Likelihood:** probability that event in C will have observable X
- $P(x_1,x_2|C=1)$ is the probability that a high-risk customer has his $X_1=x_1$,$X_2=x_2$

Bayes Rule

prior *likelihood*

posterior

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

evidence

- Evidence: P(x) probability that observation x is seen regardless if positive or negative

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

Bayes' Rule

prior *likelihood*

posterior

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

evidence

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

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

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

Bayes Rule for classification

- Assume know : prior, evidence and likelihood
- Will learn how to estimate them from the data later
- Plug them in into Bayes formula to obtain P(C|x)
- Choose C=1 if P(C=1|x)>P(c=0|x)

Bayes Rule for classification

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Bayes' Rule: K>2 Classes

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

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

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

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

Bayes' Rule

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

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

- Deciding on specific input x
- P(x) is the same for all classes
- Don't need it to compare posterior

Losses and Risks

- Decisions/Errors are not equally good or costly
- e.g an accepted low-risk applicant in increases profit, while a rejected high-risk decreases loss.
- However, the loss for a high-risk applicant accepted can be different from loss from incorrectly rejecting low-risk apllicant
- What about other domains like medical diagnosis or earthquake prediction?

Losses and Risks

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

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

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

Losses and Risks: 0/1 Loss

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

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

For minimum risk, choose the most probable class

Losses and Risks: Reject

- In some applications, wrong decisions (misclassification have high cost)
- Manual decision is made if the system has low uncertainty
- An additional action *reject* or *doubt* is added.

Losses and Risks: Reject

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

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

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

Losses and Risks: Reject

The optimal decision rule is to

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

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

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

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

- Define a function $g_i(x)$ for each class ("goodness" of selecting class C_i given observables x)

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

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

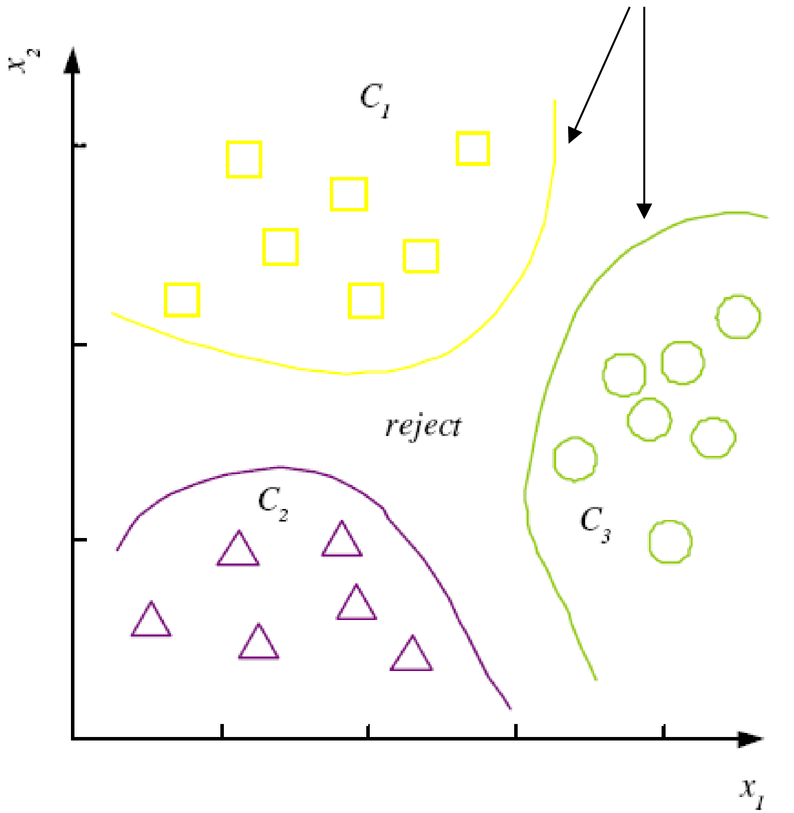
- Maximum discriminant corresponds to minimum conditional risk

Decision Regions

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

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

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



K=2 Classes

- $g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$

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

- *Log odds:*

$$\log \frac{P(C_1 | \boldsymbol{x})}{P(C_2 | \boldsymbol{x})}$$

Association Rules

- Association rule: $X \rightarrow Y$

  X is called the antecedent

  Y is called the consequent
- People who buy X typically also buy Y
- If there is a customer who buy X and does not buy Y, he is a potential Y customer

Association Rules

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

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

- **Confidence** ($X \rightarrow Y$):

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

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

Association Rules

Transaction number	Items
0	soy milk, lettuce
1	lettuce, diapers, wine, chard
2	soy milk, diapers, wine, orange juice
3	lettuce, soy milk, diapers, wine
4	lettuce, soy milk, diapers, orange juice

Calculate support for {soy milk,diapers}
Calculate confidence for {diapers->wine}
Find all the set of items with support greater than 0.5 How to do that?

An example

t1:	Beef, Chicken, Milk
t2:	Beef, Cheese
t3:	Cheese, Boots
t4:	Beef, Chicken, Cheese
t5:	Beef, Chicken, Clothes, Cheese, Milk
t6:	Chicken, Clothes, Milk
t7:	Chicken, Milk, Clothes

- Transaction data
- Assume:

minsup = 0.3
minconf = 0.8%

- An example frequent *itemset*:

{Chicken, Clothes, Milk} [sup = 3/7]

- Association rules from the itemset:

Clothes $\longrightarrow$ Milk, Chicken [sup = 3/7, conf = 3/3]

… …

Clothes, Chicken $\longrightarrow$ Milk, [sup = 3/7, conf = 3/3]

Taken from: Bing Liu UIC

Association Rule

- Only one customer bought chips
- Same customer bought beer
- $P(C|B)=1$
- But support is tiny
- Support shows statistical significance

Finding Association Rules

- Step 1: Finding frequent item sets, those which have enough support
- Step 2: Converting them to rules with enough confidence

2^n

-

Step 1: A priori principle

- Suppose that we have 4 products {0,1,2,3},
- How to calculate the support for a given set.
  - Go to every transaction, check if {0,3} is present then divide by the number of transactions 2^n
- What are the possible combinations of items?
-

A priori principle

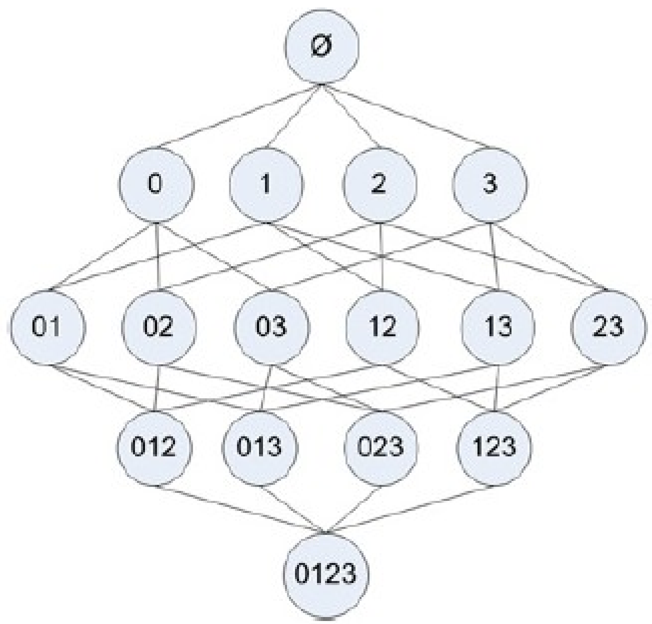
- Suppose that we have 4 products {0,1,2,3},
- How to calculate the support for a given set.
  - Go to every transaction, check if {0,3} is present then divide 2^n by the number of transactions
- what are the possible combinations of items?

$$2^n$$

Only 100 items will generate $1.26*10^{30}$ possibilities.

Taken from the book Machine Learning in Action

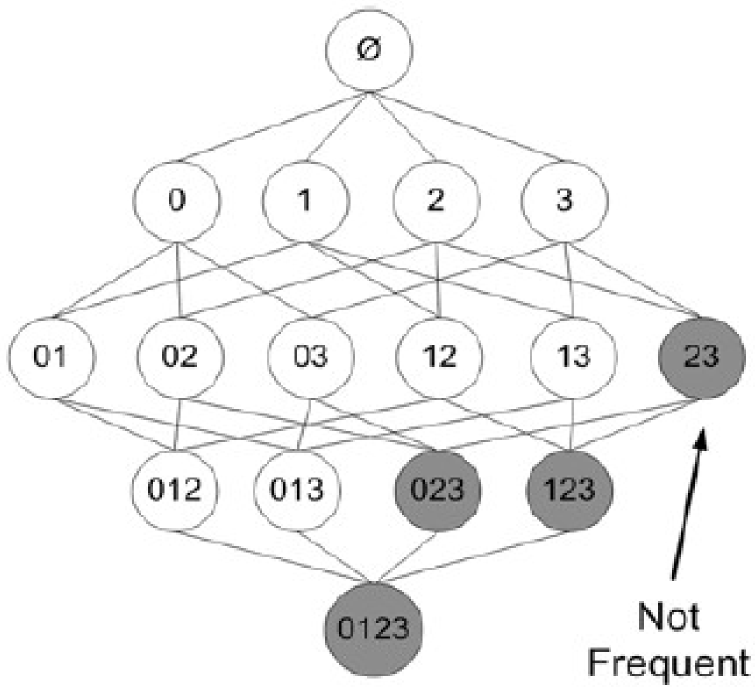
Step 1: A priori principle



If an item set is frequent, all its subsets are frequent

Taken from the book Machine Learning in Action

A priori principle



If a subset is infrequent, the set is infrequent

Taken from the book Machine Learning in Action

Dataset T

minsup=0.5

TID	Items
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

1: Finding frequent itemsets

itemset:count

1. scan T ➔ C_1: {1}:2, {2}:3, {3}:3, {4}:1, {5}:3

➔ F_1: {1}:2, {2}:3, {3}:3, {5}:3

➔ C_2: {1,2}, {1,3}, {1,5}, {2,3}, {2,5}, {3,5}

2. scan T ➔ C_2: {1,2}:1, {1,3}:2, {1,5}:1, {2,3}:2, {2,5}:3, {3,5}:2

➔ F_2: **{1,3}**:2, **{2,3}**:2, **{2,5}**:3, **{3,5}**:2

➔ C_3: {2, 3,5}

3. scan T ➔ C_3: **{2, 3, 5}**:2 ➔ F_3: **{2, 3, 5}**

Example taken from: http://www2.cs.uic.edu/~liub

Step 2: Generating rules from frequent itemsets

- Frequent itemsets ≠ association rules
- One more step is needed to generate association rules
- For each frequent itemset *X*,

For each proper nonempty subset *A* of *X*,

- Let *B* = X - *A*
- A ⟶ B is an association rule if
  - Confidence(A ⟶ B) ≥ minconf,

    confidence(A ⟶ B) = support(A,B) / support(A)

Example taken from: http://www2.cs.uic.edu/~liub

Generating rules: an example

- Suppose {2,3,4} is frequent, with sup=50%
  - Proper nonempty subsets: {2,3}, {2,4}, {3,4}, {2}, {3}, {4}, with sup=50%, 50%, 75%, 75%, 75%, 75% respectively
  - These generate these association rules:
    - 2,3 $\rightarrow$ 4, confidence=100%
    - 2,4 $\rightarrow$ 3, confidence=100%
    - 3,4 $\rightarrow$ 2, confidence=67%
    - 2 $\rightarrow$ 3,4, confidence=67%
    - 3 $\rightarrow$ 2,4, confidence=67%
    - 4 $\rightarrow$ 2,3, confidence=67%
    - All rules have support = 50%

Example taken from: http://www2.cs.uic.edu/~liub

Generating rules: summary

- To recap, in order to obtain A $\longrightarrow$ B, we need to have support(A,B) and support(A)
- All the required information for confidence computation has already been recorded in itemset generation. No need to see the data *T* any more.
- This step is not as time-consuming as frequent itemsets generation.