Principles of Machine Learning

CSCI-B455

Bayesian Decision Theory

M. Oğuzhan Kulekci

Review of Basic Elements in Probability

- Random experiment, uncertainty
- Sample space S, S is either discrete or continuous
- Event E is a subset of S, which appears with probability $P(E)$
- $P(E)$: Repeat an experiment many times, count how many times outcome is E.
- $0 \leq P(E) \leq 1, \; P(S) = 1$
- E_i and E_j are mutually exclusive if $E_i \cap E_j = \emptyset$
- For exclusive events $P(\cup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$
- If $E \cap F \neq \emptyset$, then $P(E \cup F) = P(E) + P(F) - P(F \cap U)$
- E^c is the complement of event E and $P(E) + P(E^c) = 1$

Review of Basic Elements in Probability

- **Conditional Probability** $P(E\,|\,F)$: Probability of E given that F has occurred, $P(E\,|\,F) = \dfrac{P(E \cap F)}{P(F)}$
- $P(F\,|\,E) = \dfrac{P(F \cap E)}{P(E)}$, $P(E \cap F) = P(F \cap E) \Rightarrow P(E\,|\,F) \cdot P(F) = P(F\,|\,E) \cdot P(E)$
- Bayes' formula $P(E\,|\,F) = \dfrac{P(E \cap F)}{P(F)}$
- Mutually exclusive and exhaustive events: $\cup_{i=1}^{k} F_i = S$. Then $P(E) = \displaystyle\sum_{i=1}^{n} P(E \cap F_i) = \sum_{i=1}^{n} P(E\,|\,F_i) \cdot P(F_i)$
- $P(F_i\,|\,E) = \dfrac{P(F_i \cap E)}{P(E)} = \dfrac{P(E\,|\,F_i) \cdot P(F_i)}{\sum_{i=1}^{n} P(E\,|\,F_i) \cdot P(F_i)}$

Bayes' example

- There is a disease that appears with probability one in a million. There is a test that can detect the disease with probability 99% on a person with the disease. However, with 1/1000 probability, the test reports positive on a healthy person. What is the probability that a patient is sick when the test result is positive?

$$\left.\boxed{\begin{array}{l} P(d=1)=1/10^6,\ P(d=0)=1-1/10^6, \\ P(t=1\,|\,d=1)=0.99, P(t=1\,|\,d=0)=0.001 \end{array}}\right\}\ P(d=1\,|\,t=1)=\,?$$

$$P(d=1\,|\,t=1)=\frac{P(t=1\,|\,d=1)\cdot P(d=1)}{P(t=1)}=\frac{\overset{0.99\cdot 0.000001}{P(t=1\,|\,d=1)\cdot P(d=1)}}{\underset{0.99\cdot 0.000001+0.001\cdot 0.999999}{P(t=1\,|\,d=1)\cdot P(d=1)+P(t=1\,|\,d=0)\cdot P(d=0)}}$$

$$\approx 1/10000$$

Review of Basic Elements in Probability

- **Random variable** X takes on different values based on the outcome of a random event.
- Probability distribution function of $F(a) = P(X \leq a)$, for a real number a
- $F(a) = \sum_{\forall x \leq a} P(x)$, if X is discrete, or $F(a) = \int_{-\infty}^{a} P(x)\, dx$, if X is continuous
- Joint distribution $F(x, y) = P(X \leq x, Y \leq y)$, $P(X = x) = \sum_{j} P(x, y_j)$ or $P(X = x) = \int_{-\infty}^{\infty} P(x, y)\, dy$
- Conditional distribution $P(X = x \mid Y = y) = \dfrac{P(x, y)}{P(y)}$

Review of Basic Elements in Probability

- **Expected value of a random variable X : $E[X] = \begin{cases} \sum_i x_i P(x_i) & \text{if } X \text{ is discrete} \\ \int x P(x) dx & \text{if } X \text{ is continuous} \end{cases}$**

- $E[aX + b] = aE[X] + b$; $E[X + Y] = E[X] + E[Y]$

- $E[g(X)] = \begin{cases} \sum_i g(x_i) P(x_i) & \text{if } X \text{ is discrete} \\ \int g(x) P(x) dx & \text{if } X \text{ is continuous} \end{cases}$

- **The nth moment of X: $E[X^n] = \begin{cases} \sum_i x_i^n P(x_i) & \text{if } X \text{ is discrete} \\ \int x^n P(x) dx & \text{if } X \text{ is continuous} \end{cases}$**

- First moment is $\mu = E[X]$

Review of Basic Elements in Probability

- **Variance of a random variable** X **measures expected variation of** X **around** $\mu = E[X]$

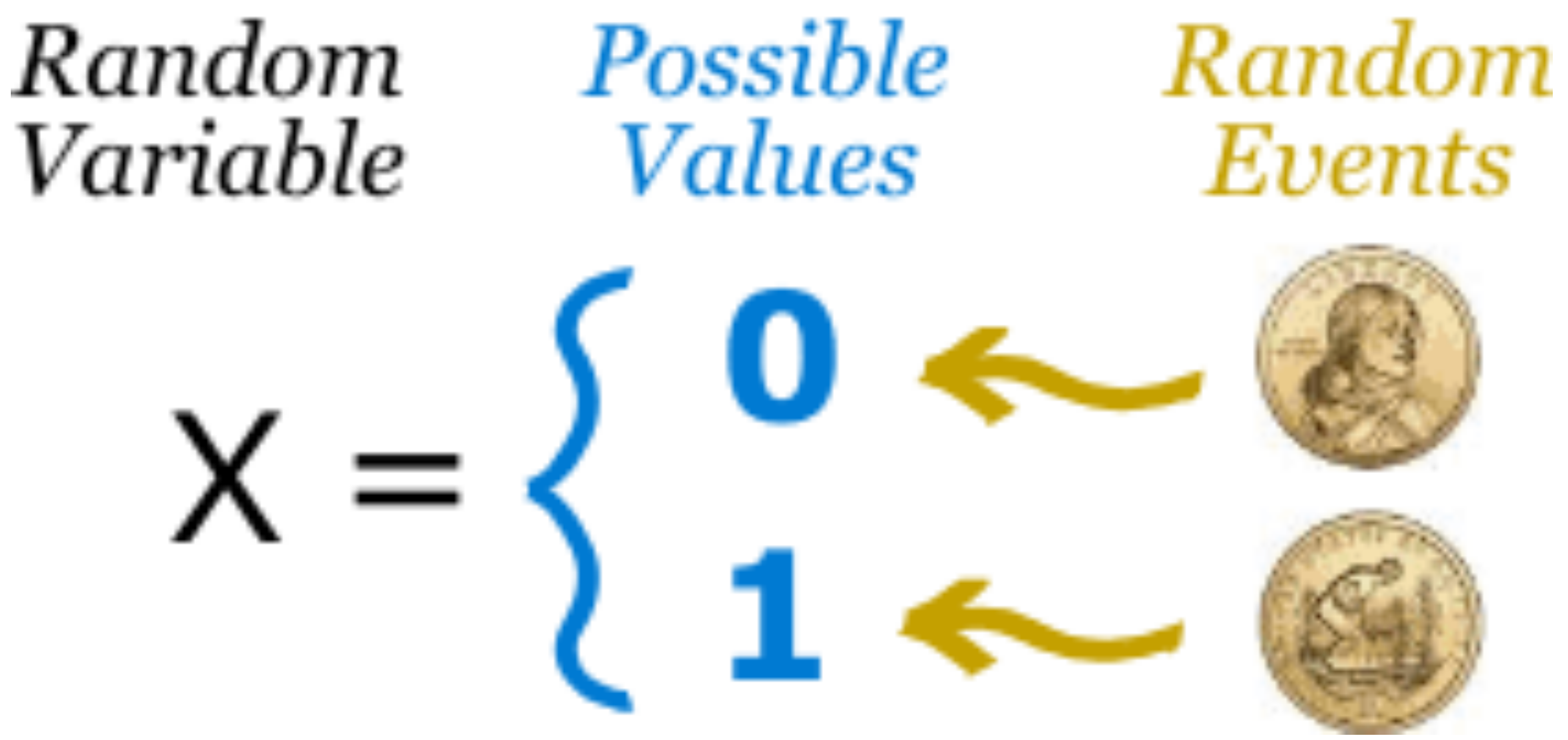
$$Var(X) = E[(X-\mu)^2] = E[X^2 - 2X\mu + \mu^2] = E[X^2] - 2\mu^2 + \mu^2 = \mathbf{E[X^2]} - \mu^{\mathbf{2}} = \sigma^2$$

- **Standard deviation** $\sigma = \sqrt{Var(X)}$

- **Covariance is the relationship between two random variables** $Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$

- **Correlation is the normalized covariance** $Cor(X, Y) = \dfrac{Cov(X, Y)}{\sqrt{Var(x)Var(Y)}}$, $-1 \leq Corr(X, Y) \leq 1$

Random Variable & Random Process



https://www.mathsisfun.com/data/random-variables.html

- TOSSING A COIN
- The outcome depends on some **unobservable** parameters (the physics of tossing it).
- It is a **random event** as the outcome, is uncertain.
- The outcome is indeed observable, and is a **random variable**, say X.
- $X = 0$ (say heads) with some probability p_0, and $X = 1$(tails) with probability $p_1 = 1 - p_0$.

- If p_0 (and hence p_1) is known, what would be a good strategy to predict the next ?
- Choose the more probable one, to reduce the **error** ,(1-selected.probability).

Random Variable & Random Process

- If more probable one is **not known**, but we have a sequence of previous outcomes, how do we proceed ?

We need an estimator, which is easy $\hat{p}_0 = \dfrac{\textit{number of head tosses}}{\textit{total number of tosses}}$.

Assume, previous outcomes are H,T,T,T,H,H,T,T,T,H. Then $\hat{P}(Heads) = \hat{p}_0 = \dfrac{4}{10}$

Based on this, the prediction next will be *tails.*

Notice that, after the next toss, we should update our approximation!

Classification with probabilities

- Revisiting the credit scoring, low-risk and high-risk customers
- The parameters we use X_1, X_2 as the yearly income and savings.
- When a new customer arrives, the bank wants to predict the credit score class.
- The class label of the customer, $C,$ is a **Bernoulli** random variable

- ***Bernoulli*** *random variable* X *has two outcomes, Success (1) or Fail (0).*
- *Success probability shown by* $P(Success) = P(1) = p$
- *Failure probability is* $(1 - p)$
- $E(X) = p$ *and* $Var(X) = \sigma^2 = (1-p)^2 \cdot p + (0-p)^2 \cdot (1-p) = p(1-p)$

Classification with probabilities

Choose the class with higher probability for the given input (x_1, x_2)

$$C = 1, \, if \; P(C = 1 \,|\, x_1, x_2) > P(C = 0 \,|\, x_1, x_2), \text{ else } C = 0$$

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

Same story with the coin tossing, but decision depends on two variables.

How to compute $P(C \,|\, x_1, x_2)$ by using the training data ?

Classification with probabilities

- **Prior Probability:** $P(C=1)$, regardless of the input.
- **Class likelihood:** $P(\mathbf{x} \mid C)$, the probability of $\mathbf{x}$ in class C
- **Evidence:** $P(\mathbf{x})$, the probability of observing $\mathbf{x}$ regardless of the class.
- **Posterior Probability:** $P(C \mid \mathbf{x})$, given input, predict its class, **the aim.**

$$posterior = \frac{prior \cdot likelihood}{evidence} = \frac{P(C) \cdot P(\mathbf{x} \mid C)}{P(\mathbf{x})} = P(C \mid \mathbf{x})$$

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

Classification with probabilities

- **Prior Probability:** $P(C=1)$, regardless of the input.
- **Class likelihood:** $P(\mathbf{x}\,|\,C)$, the probability of $\mathbf{x}$ in class C
- **Evidence:** $P(\mathbf{x})$, the probable,lity of observing $\mathbf{x}$ regardless of the class.
- **Posterior Probability:** $P(C\,|\,\mathbf{x})$, given input, predict its class, **the aim.**

$$posterior = \frac{prior \cdot likelihood}{evidence} = \frac{P(C) \cdot P(\mathbf{x}\,|\,C)}{P(\mathbf{x})} = P(C\,|\,\mathbf{x})$$

- **In case of multiple classes** $C_1, C_2, \ldots, C_K$**:**

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

Losses and Risks

- **Wrong decisions (misclassifications) might not be equally costly.**
- Compare the consequences of wrong decision on low-risk and high-risk customers.
- We need an adjustment while calculating the error.

- α_i : classifying input in class C_i
- $\lambda_{i,j}$: incurred loss when input is classified as C_i, while it is C_j .

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{K} \lambda_{i,j} \cdot P(C_j \mid \mathbf{x})$$

Choose C_i when $R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$

Losses and Risks

An example case: Given $\lambda_{1,1} = 0, \lambda_{2,2} = 0, \lambda_{1,2} = 10, \lambda_{2,1} = 5$, what would be the optimal decision rule?

Risk when we decide on C_1 :

$$R(\alpha_1 | x) = \lambda_{1,1} \cdot P(C_1 | x) + \lambda_{1,2} \cdot \mathbf{P(C_2 | x)} = 0 \cdot P(C_1 | x) + 10 \cdot (\mathbf{1 - P(C_1 | x)})$$

Risk when we decide on C_2 :

$$R(\alpha_2 | x) = \lambda_{2,1} \cdot P(C_1 | x) + \lambda_{2,2} \cdot P(C_2 | x) = 5 \cdot P(C_1 | x) + 0 \cdot P(C_2 | x) = 5P(C_1 | x)$$

Choose C_1 when $R(\alpha_1 | x) < R(\alpha_2 | x)$, which means

$$10 \cdot (1 - P(C_1 | x)) < 5 \cdot P(C_1 | x) \Rightarrow 2 - 2P(C_1 | x) < P(C_1 | x) \Rightarrow \boxed{\frac{2}{3} < P(C_1 | x)}$$

Losses and Risks

How do we define the case with equal costs of misclassification?

- **Assume $\lambda_{i,j} = 0$ if $i = j$, and $\lambda_{i,j} = 1$ for all $i \neq j$. Then:**

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

Therefore, choosing the maximum posterior probability, guarantees the minimum risk, when misclassification costs are equal.

Notice that this is actually very rare in practice!

Losses and Risks

How do we define the case with equal costs of misclassification?

If $\lambda_{1,1} = 0, \lambda_{2,2} = 0, \lambda_{1,2} = \lambda_{2,1} = w$, what would be the optimal decision rule?

Risk when we decide on C_1 :

$$R(\alpha_1 | x) = \lambda_{1,1} \cdot P(C_1 | x) + \lambda_{1,2} \cdot P(C_2 | x) = 0 \cdot P(C_1 | x) + w \cdot (1 - P(C_1 | x))$$

Risk when we decide on C_2 :

$$R(\alpha_2 | x) = \lambda_{2,1} \cdot P(C_1 | x) + \lambda_{2,2} \cdot P(C_2 | x) = w \cdot P(C_1 | x) + 0 \cdot P(C_2 | x) = wP(C_1 | x)$$

Choose C_1 when $R(\alpha_1 | x) < R(\alpha_2 | x)$, which means

$$w \cdot (1 - P(C_1 | x)) < w \cdot P(C_1 | x) \Rightarrow 1 - P(C_1 | x) < P(C_1 | x) \Rightarrow \boxed{\frac{1}{2} < P(C_1 | x)}$$

Losses and Risks

What if the cost of misclassification is extremely high ?

- When the computed risk is greater than a threshold, more complex systems will handle the input, even maybe manual.
- We introduce an additional decision α_{K+1} as **REJECT**.

The loss function $\lambda_{i,k}$ is then

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

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

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

$$1 - P(C_i | \mathbf{x}) < \lambda \rightarrow P(C_i | \mathbf{x}) > 1 - \lambda$$

Choose C_i if $R(\alpha_i | \mathbf{x}) < R(\alpha_j | \mathbf{x}), for\ i = 1,2,..,K, K+1$,which means $P(C_i | \mathbf{x}) > 1 - \lambda$

Losses and Risks

What if the cost of misclassification is extremely high ?

If $\lambda_{1,1} = 0,\ \lambda_{2,2} = 0,\ \lambda_{1,2} = 10,\ \lambda_{2,1} = 5$ and $\lambda_{r,1} = \lambda_{r,2} = 1$, what would be the optimal decision rule?

Risk when we decide on C_1 :$R(\alpha_1 | x) = \lambda_{1,1} \cdot P(C_1 | x) + \lambda_{1,2} \cdot \mathbf{P}(\mathbf{C_2} | \mathbf{x}) = 0 \cdot P(C_1 | x) + 10 \cdot (\mathbf{1} - \mathbf{P}(\mathbf{C_1} | \mathbf{x}))$

Risk when we decide on C_2 :$R(\alpha_2 | x) = \lambda_{2,1} \cdot P(C_1 | x) + \lambda_{2,2} \cdot P(C_2 | x) = 5 \cdot P(C_1 | x) + 0 \cdot P(C_2 | x) = 5P(C_1 | x)$

Risk when we decide on *reject* :$R(\alpha_r | x) = \lambda_{r,1} \cdot P(C_1 | x) + \lambda_{r,2} \cdot P(C_2 | x) = 1 \cdot P(C_1 | x) + 1 \cdot (1 - P(C_1 | x) = 1$

To choose C_1, $R(\alpha_1 | x) < R(\alpha_r | x)$ $10 - 10P(C_1 | x) < 1 \Rightarrow P(C_1 | x) > \frac{9}{10}$

To choose C_2, $R(\alpha_2 | x) < R(\alpha_r | x)$ $5 \cdot P(C_1 | x) < 1 \Rightarrow P(C_1 | x) < \frac{1}{5}$

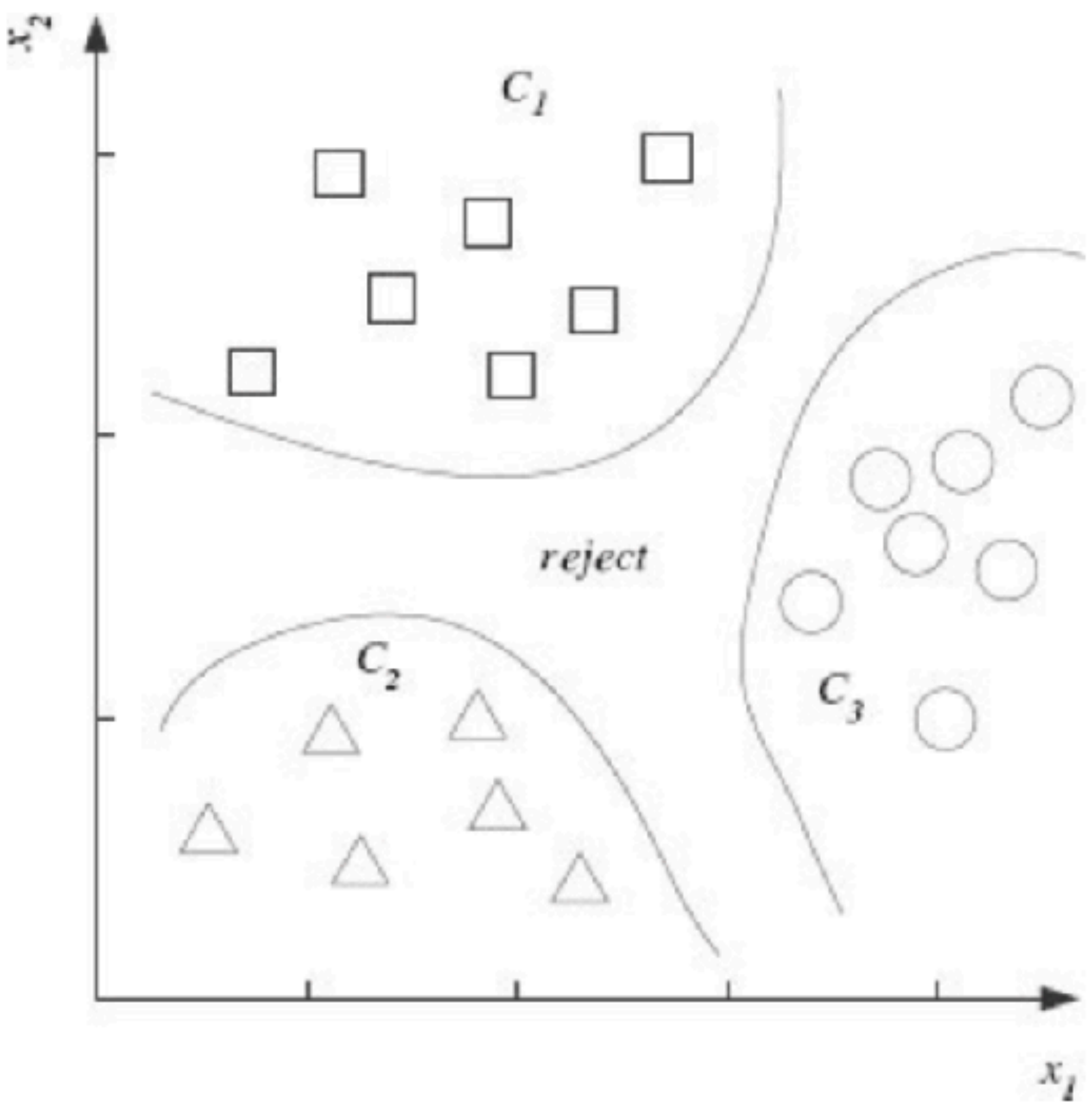
Discriminant Functions

Discriminant functions can be used as classifiers : Choose class C_i, if $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$, for $g_i(\mathbf{x}) = -R(\alpha_i | \mathbf{x})$

Assuming 0/1 *loss* function, $g_i(\mathbf{x}) = -R(\alpha_i | \mathbf{x}) = P(C_i | \mathbf{x})$

Even by neglecting the common denominator $p(\mathbf{x})$,

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



Discriminant Functions

An example: Let the likelihood ratio be $\ell = \dfrac{P(x \mid C_1)}{P(x \mid C_2)}$.

If we define discriminant function as $g(x) = \dfrac{P(C_1 \mid x)}{P(C_2 \mid x)}$

$$g(x) = \frac{P(C_1 \mid x)}{P(C_2 \mid x)} = \frac{P(x \mid C_1) \cdot P(C_1)/P(x)}{P(x \mid C_2) \cdot P(C_2)} = \frac{P(x \mid C_1) \cdot P(C_1)/P(x)}{P(x \mid C_2) \cdot P(C_2)} = \ell \cdot \frac{P(C_1)}{P(C_2)}.$$

Notice that if $P(C_1) = P(C_2)$, likelihood ratio becomes directly the discriminant.

Reading Material

- Chapter 3, excluding 3.5 Association Rules