SYDE 372
Introduction to Pattern Recognition

Probability Measures for Classification: Part I

Alexander Wong

Department of Systems Design Engineering
University of Waterloo

Outline

Why use probability measures for classification?

- Great variability may occur within a class of patterns due to measurement noise (e.g., image noise and warping) and inherent variability (apples can vary in size and shape)
- We tried to account for these variabilities by treating patterns as random vectors
- In the MICD classifier, we account for this variability by incorporating statistical parameters of the class (e.g., mean and variance)
- This works well for scenarios where class distributions can be well modeled based on Gaussian statistics, but may perform poorly when the class distributions are more complex and non-Gaussian.
- How do we deal with this?

Why use probability measures for classification?

- Idea: What if we have more complete information about the probabilistic behaviour of the class?
- Given known class conditional probability density distributions, we can create powerful similarity measures that tell us the **likelihood**, or **probability**, of each class given an observed pattern.
- Classifiers built on such probabilistic measures are optimal in the minimum **probability of error** sense.

Bayesian classifier

- Consider the two class pattern recognition problem:
  - Given an unknown pattern $\underline{x}$, assign the pattern to either class A or class B.
- A general rule of statistical decision theory is to minimize the "cost" associated with making a wrong decision.
  - e.g., amount of money lost by deciding to buying a stock that gets delisted the next day and is actually a "don't buy".

Bayesian classifier

- Let L_{ij} be the cost of deciding on class c_j when the true class is c_i
- The total risk associated with deciding $\underline{x}$ belongs to c_j can be defined by the expected cost:

$$r_j(\underline{x}) = \sum_{i=1}^{K} L_{ij} P(c_i|\underline{x}) \tag{1}$$

where K is the number of classes and $P(c_i|\underline{x})$ is the posterior distribution of class c_i given the pattern $\underline{x}$.

Bayesian classifier

- For the two class case:

$$r_1(\underline{x}) = L_{11}P(c_1|\underline{x}) + L_{21}P(c_2|\underline{x}) \tag{2}$$

$$r_2(\underline{x}) = L_{12}P(c_1|\underline{x}) + L_{22}P(c_2|\underline{x}) \tag{3}$$

- Applying Bayes' rule gives us:

$$r_1(\underline{x}) = \frac{L_{11}P(\underline{x}|c_1)P(c_1) + L_{21}P(\underline{x}|c_2)P(c_2)}{p(\underline{x})} \tag{4}$$

$$r_2(\underline{x}) = \frac{L_{12}P(\underline{x}|c_1)P(c_1) + L_{22}P(\underline{x}|c_2)P(c_2)}{p(\underline{x})} \tag{5}$$

Bayesian classifier

- The general K-class Bayesian classifier is defined as follows, and minimizes total risk:

$$\underline{x} \in c_i \text{ iff } r_i(\underline{x}) < r_j(\underline{x}) \;\; \forall\, j \neq i \tag{6}$$

- For the two class case:

$$(L_{11} - L_{12})P(\underline{x}|c_1)P(c_1) \underset{2}{\overset{1}{\gtrless}} (L_{21} - L_{22})P(\underline{x}|c_2)P(c_2) \tag{7}$$

- How do you choose an appropriate cost?

Choosing cost functions

- The most common cost used in the situation where no other cost criterion is known is the “zero-one” loss function:

$$L_{ij} = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \tag{8}$$

- Meaning: all errors have equal costs.
- Given the “zero-one” loss function, the total risk function becomes:

$$r_j(\underline{x}) = \sum_{\substack{i=1 \\ i \neq j}}^{K} P(c_i|\underline{x}) = P(\epsilon|\underline{x}) \tag{9}$$

- So minimizing total risk in this case is the same as minimizing probability of error!

Types of probabilistic classifiers

- Using the “zero-one” loss function, we will study two main types of probabilistic classifiers:
  - Maximum a Posteriori (MAP) probability classifier
  - Maximum Likelihood (ML) classifier

Maximum a Posteriori classifier

- Given two classes A and B, the MAP classifier can be defined as follows:

$$P(A|\underline{x}) \underset{B}{\overset{A}{\gtrless}} P(B|\underline{x}) \tag{10}$$

  where $P(A|\underline{x})$ and $P(B|\underline{x})$ are the posterior class probabilities of A and B, respectively, given observation $\underline{x}$.
- **Meaning:** All patterns with a higher posterior probability for A than for B will be classified as A, and all patterns with a higher posterior probability for B than for A will be classified as B

Maximum a Posteriori classifier

- Class probability models usually given in terms of class conditional probabilities $P(\underline{x}|A)$ and $P(\underline{x}|A)$
- More convenient to express MAP in the form:

$$\frac{P(\underline{x}|A)}{P(\underline{x}|B)} \underset{B}{\overset{A}{\gtrless}} \frac{P(B)}{P(A)} \tag{11}$$

$$l(\underline{x}) \underset{B}{\overset{A}{\gtrless}} \theta \tag{12}$$

where $l(\underline{x})$ is the likelihood ratio and θ is the threshold

Maximum a Posteriori classifier

- When dealing with probability density functions with exponential dependence (e.g., Gamma, Gaussian, etc.), it is more convenient to deal with MAP in the log-likelihood form:

$$\log l(\underline{x}) \underset{B}{\overset{A}{\gtrless}} \log \theta \tag{13}$$

Maximum a Posteriori classifier: Example

- Suppose we are given a two-class problem, where $P(x|A)$ and $P(x|B)$ are given by:

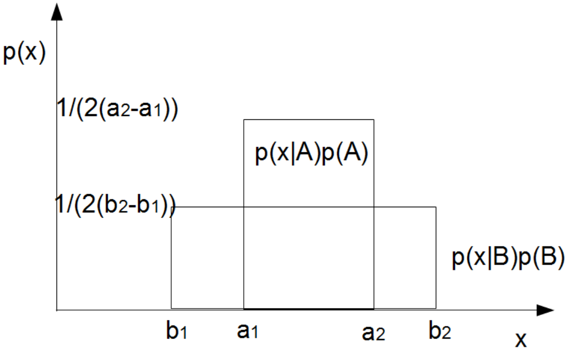
$$p(x|A) = \begin{cases} 0, & x < a_1 \\ \frac{1}{a_2 - a_1} & a_1 \leq x \leq a_2 \\ 0 & a_2 < x \end{cases} \tag{14}$$

$$p(x|B) = \begin{cases} 0, & x < b_1 \\ \frac{1}{b_2 - b_1} & b_1 \leq x \leq b_2 \\ 0 & b_2 < x \end{cases} \tag{15}$$

where $b_2 > a_2 > a_1 > b_1$.

- Assuming $P(A) = P(B) = 1/2$, develop the MAP classification strategy.

Maximum a Posteriori classifier: Example



Maximum a Posteriori classifier: Example

- The MAP classification strategy can be defined as:
  - $b_1 < x < a_1$: Decide class B
  - $a_1 < x < a_2$: Decide class A
  - $a_2 < x < b_2$: Decide class B
  - Otherwise: No decision

Maximum a Posteriori classifier

- When dealing with probability density functions with exponential dependence (e.g., Gamma, Gaussian, etc.), it is more convenient to deal with MAP in the log-likelihood form:

$$\log l(\underline{x}) \underset{B}{\overset{A}{\gtrless}} \log \theta \tag{16}$$

Maximum Likelihood classifier

- Ideally, we would like to use the MAP classifier, which chooses the most probable class:

$$\frac{P(\underline{x}|A)}{P(\underline{x}|B)} \underset{B}{\overset{A}{\gtrless}} \frac{P(B)}{P(A)} \tag{17}$$

- However, in many cases the priors $P(A)$ and $P(B)$ are unknown, making it impossible to use the posteriors $P(A|\underline{x})$ and $P(B|\underline{x})$.
- Common alternative is, instead of choosing the most probable class, we choose the class that makes the observed pattern $\underline{x}$ most probable.

Maximum Likelihood classifier

- Instead of maximizing the posterior, we instead maximize the likelihood:

$$P(\underline{x}|A) \underset{B}{\overset{A}{\gtrless}} P(\underline{x}|B) \tag{18}$$

- In likelihood form:

$$\frac{P(\underline{x}|A)}{P(\underline{x}|B)} \underset{B}{\overset{A}{\gtrless}} 1 \tag{19}$$

- Can be viewed as special case of MAP where $P(A) = P(B)$.