Lecture 4
Basic Statistic Learning

Topics of this lecture

- Decision making based on probabilities
  - Likelihood based decision making
  - Joint probability based decision making
  - Posterior probability based decision making
- Probability density function estimation
  - Parametric approaches
  - Non-parametric approaches
- Naïve Bayes classifier

What is likelihood?

- Consider a pattern x generated with a probability controlled by some hidden factor θ.
- Given θ, the ***conditional probability*** of x (e.g. when the value of the random variable X equals to x) is denoted by $p(x|\theta)$.
- On the other hand, given x, the ***likelihood*** that this x is controlled by θ is $L(\theta|x) = p(x|\theta)$.

What is likelihood?

- Example: In a coin tossing game, if we observe "x=HH", the likelihood that the coin is fair (head and tail occur with the same probability) is given by $L(\theta = 0.5|x = HH) = p(x = HH|\theta = 0.5) = 0.5^2 = 0.25$.
- For this example, we may also assume that the probability to get a head is $\theta = 0.6$, and the likelihood is $p(x = HH|\theta = 0.6) = 0.6^2 = 0.36$. That is, when "$x = HH$" is observed, the coin is more likely to be un-fair.
- If we observe "$x = HHHHH$", it is more likely that the coin is not fair.
- That is, likelihood provides a way to measure how likely an assumption is true when a concrete observation is given.

Decision based on likelihood

- Let us consider a pattern classification problem with N classes $C_1, C_2, \dots, C_N$.
- For a newly observed pattern x, we want to determine which class x belongs to.
- We can consider that each class has a **hidden factor** for controlling a random process for generating x. That is, x can be generated with a conditional probability $p(x|C_i)$.
- Given a pattern x, we may classify it to the i-th class if

$$i = arg \max_j p(x|C_j)$$

- The rationale is that, if x belongs to C_i, x can occur with the highest probability. In other words, the hidden factor is most likely provided by C_i.

Decision based on joint probability

- Decisions based on likelihood may not be good in practice. For example, for a two class problem, patterns in C_1 may occur more frequently than those in C_2.
- To reduce the total number of mistakes in decision making, it is better to use $p(C_i)$ as a weight, and assign x to C_i if

$$i = arg\max_j p\left(x\middle|C_j\right)p(C_j)$$

- Note that $p\left(x\middle|C_j\right)p\left(C_j\right) = p(x, C_j)$ is the joint probability.
- By marginalizing the joint probability, we can find $p(x)$, which can be used to generate new data.
- Thus, joint probability-based decision making is also ***generative***.

Decision based on posterior probability

- In practice, if we just want to assign a given pattern x to some class, it is not necessary to know the joint probability. Instead, we may use the posterior probability $p(C_i|x)$.
- That is, a given pattern x can be assigned to the i-th class if

$$i = arg\max_j p(C_j|x)$$
$$= arg\max_j \frac{p\left(x\middle|C_j\right)p(C_j)}{\sum_{j=1}^{N} p\left(x\middle|C_j\right)p(C_j)}$$

- Here, we have used the well-known Bayesian theorem. Since the denominator is common for all classes, decision based on the posterior probability is the same as the one based on joint probability.

Minimum mistake decision

- Let us define the loss function as the "zero-one loss function" given by

$$l(y = i \mid x \in C_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

- Here, we assign no loss to a correct classification and a unit loss to a misclassification. Also, we assume that all mistakes are equally costly.
- Given an observation x, the risk is found by

$$R(y = i|x) = \sum_{j=1}^{N} l\left(y = i\middle|x \in C_j\right)p\left(C_j|x\right) = \sum_{j \neq i}^{N} p(C_j|x) = 1 - p(C_i|x)$$

- If we assign x to the class with the maximum posterior probability, the risk of misclassification can be minimized.

Statistic approaches and Deterministic approaches (1)

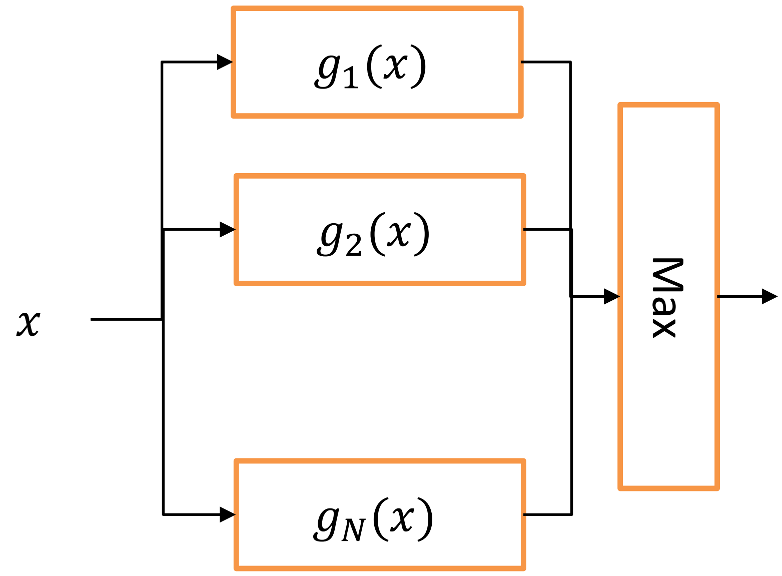
- Generally speaking, estimation of the posterior probability can be easier than estimation of the joint probability.
- In practice, however, estimation of the posterior probability can be difficult, too, especially in the case we do not have enough data.
- In this situation, we can find a discriminant function for each class, say $g_i(x)$, using deterministic approaches.

Statistic approaches and Deterministic approaches (2)

- For any given pattern x, it is assigned to the i-th class if

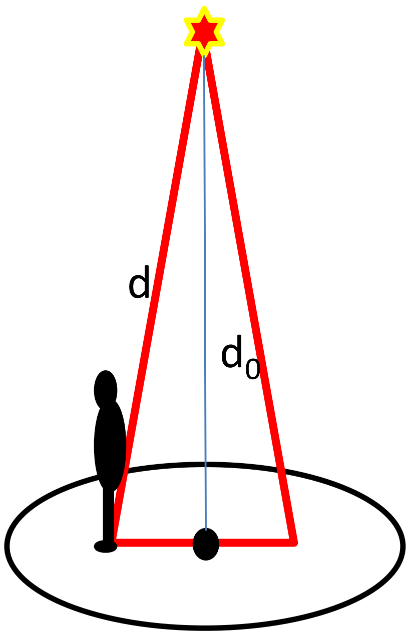
$$i = arg\max_j g_j(x)$$

- Generally speaking, neural network-based learning belongs to this kind of approaches.
- Note that the likelihood, the posterior probability, and the joint probability are also discriminant functions.



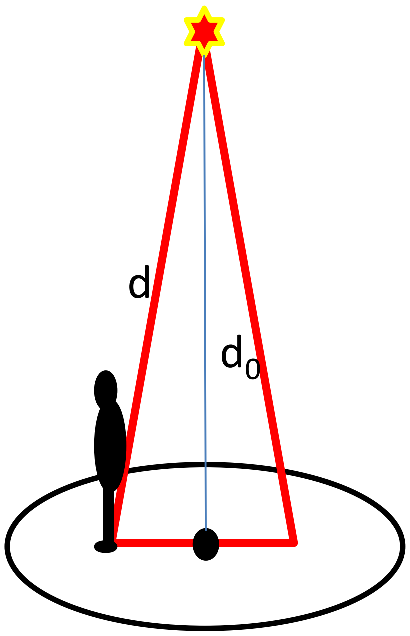
Example 1: Determine the distance between a subject and a binary sensor

- A binary infrared sensor usually outputs 1 when a person is detected, and 0 otherwise.
- We can consider the sensor an unfair coin. When other factors fixed, the ***unfairness*** depends on the distance d between the sensor and the subject; and can be defined as the conditional probability $p = p(1|d)$.
- Example: If we observe a sequence 1111110111, we may think that d is almost d_0 (i.e. the subject is just under the sensor), and the likelihood is given by $p(1111110111 \mid d_0) = {p_0}^9(1 - p_0)$.



Example 1: Determine the distance between a subject and a binary sensor

- To make a decision, we may find the likelihoods for several different d values, and choose the one with the maximum likelihood.
- If the subject visits different places with different "preferences", we can use $p(d)$ to modify the decision, and reduce the errors using the maximum posterior decision.



Probability density function estimation

- To use the statistic approaches, we must estimate the probability density function (pdf).
- There are mainly two approaches:
  - Parametric approaches and
  - Non-parametric approaches
- Parametric approaches can be more efficient if we know the "type" of the pdf, because what is needed is to estimate a few parameters that controls the pdf (e.g. mean, variance, etc.)
- Non-parametric approaches can be more flexible.

Parameter estimation for Bernoulli distribution (1)

- As in sensor data analysis, Bernoulli distribution is useful in many applications.
- Bernoulli distribution is controlled by the value μ, which is the probability that the positive event occurs (and the probability that the negative event occurs is $1-\mu$).
- Suppose that X is a random variable following the Bernoulli distribution. The probability that $X = x$ (0 or 1) is given by

$$p(x|\mu) = \mu^x (1-\mu)^{1-x}$$

- In fact, μ is the mean of the Bernoulli distribution.

Parameter estimation for Bernoulli distribution (2)

- Suppose that we observed $X = \{x_1, x_2, \ldots, x_N\}$. Each element is 0 or 1 ***independently and identically*** generated by following the Bernoulli distribution.
- We want to estimate the mean based on these data.
- The likelihood that the mean equals to μ is given by

$$p(X|\mu) = \prod_{i=1}^{N} p(x_i|\mu) = \prod_{i=1}^{N} \mu^{x_i}(1-\mu)^{1-x_i}$$

Parameter estimation for Bernoulli distribution (3)

- To simplify the problem, we take the natural log and use the following log likelihood:

$$\ln p(X|\mu) = \sum_{i=1}^{N} \ln p(x_i|\mu)$$
$$= \sum_{i=1}^{N}[x_i \ln \mu + (1 - x_i) \ln(1 - \mu)]$$

- Find the first order derivative of the likelihood function with respect to μ, and let is be zero, we can find the maximum likelihood estimation of the mean value as follows:

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Example 2: Estimate the activity using a binary sensor

- Similar to Example 1, we may fix all other factors, and allow the subject to conduct some activities near a binary sensor.
- Again, the probability to get a 1 from the sensor is controlled by the mean value μ of a Bernoulli distribution.
- Suppose that we have estimated the mean values $\mu_1, \mu_2, \ldots,$ for several different activities based on many observations.
- Given a new observation $X = \{x_1, x_2, \ldots, x_N\}$, we can determine the activity based on the likelihood function (or the log likelihood function) as follows:

$$i = arg \max_j \prod_{k=1}^{N} p(x_k | \mu_j)$$

Parameter estimation for Gaussian distribution (1)

- Gaussian distribution is also known as the normal distribution. The probability density function of a single random variable is given by

$$N(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- where μ and σ are the mean and the standard deviation, respectively. For D-dimensional case, the density function is given as follows:

$$N(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

- where $\boldsymbol{\mu}$ is a D-dimensional mean vector, and Σ is a DxD co-variance matrix.

Parameter estimation for Gaussian distribution (2)

- Given a data set $X = \{x_1, x_2, \dots, x_N\}$, the mean and the co-variance matrix can be estimated using the following equations:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \Sigma = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T$$

- Note that the mean is found directly using the maximum likelihood estimation, but the co-variance matrix has been modified to make it un-biased.
- Based on information theory, Gaussian distribution is the most natural distribution (i.e. has the maximum entropy).
- Also, according to the ***central limit theorem***, the mixture of many random variables can be approximated by a Gaussian.

Kernel density estimation (1)

- Kernel-based probability density estimation is a non-parametric approach.
- This approach is also referred to as Parzen–Rosenblatt window method.
- Let $(x_1, x_2, \ldots, x_N)$ be a univariate independent and identically distributed sample drawn from some distribution with an unknown density f.
- The kernel density estimator is given by

$$g_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x - x_i) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x - x_i}{h})$$

- where K is a kernel function (non-negative) and h>0 is the bandwidth.

Kernel density estimation (2)

- The simplest kernel function is the so called uniform function, which is defined by a rectangular window as follows:

$$K(u) = \frac{1}{2}, for\ |u| \leq 1$$

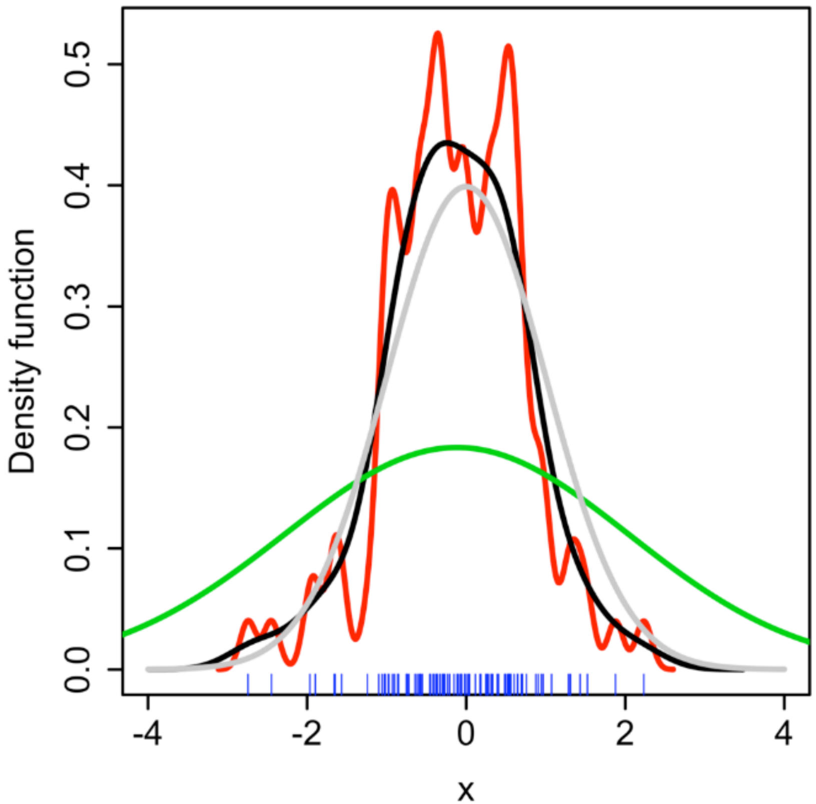
- The most popular kernel function used is the well-known Gaussian defined by

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}\mu^2\}$$

- A pdf found using the Gaussian kernel is more smooth, and can be used when the number of data is not large.
- Note that the smoothness is also controlled by the parameter h. Generally speaking, h cannot be too small, or the pdf may contain many noises; and h cannot be too large, or the pdf may not approximate the true density function well.

Kernel density estimation (3)

- Kernel density estimation with different bandwidths of a random sample of 100 points from a standard normal distribution.
  - Gray: true density (standard normal).
  - Red: h=0.05.
  - Black: h=0.337.
  - Green: h=2.



(from Wikipedia)

The Naïve Bayes classifier (1)

- Naïve Bayes Classifier (NBC) is a simple statistic learning model.
- It is simple, scalable, and useful for solving problems with very high dimensions (e.g. text classification).
- In an NBC, all elements in a feature vector are considered independent, and therefore, a high dimensional joint probability function can be found using the product of many univariate probability functions.
- Specifically, to make a decision, what we need is the posterior probability $p(C_i|x_1, x_2, \dots, x_D)$ for a given observation (a D-dimensional pattern) and the i-th class (i=1,2,…,N).

The Naïve Bayes classifier (2)

- In NBC, the posterior probability is approximated by

$$p(C_i|x_1, x_2, \ldots, x_D) \propto p(C_i, x_1, x_2, \ldots, x_D)$$
$$= p(C_i) \prod_{j=1}^{D} p(x_j|C_i)$$

- For any given D-dimensional pattern x, we can make a decision based on the posterior probability.
- In practice, the assumption that all features are independent of each other may not be true. Nevertheless, NBC has been proved useful for text analysis (e.g. spam filter).

Homework

- We may also think that the outputs of a binary sensor follows a binomial distribution if we consider number of ones in N (fixed) observed data.
- Try to re-formulate the process for determining the distance of a subject based on N observations, with the binomial distribution.