Bayesian Decision Theory

CS 550: Machine Learning

Bayesian Decision Theory

- It is the fundamental statistical approach in classification
- Here it is assumed that
  1. The decision problem is posed in probabilistic terms and
  2. All relevant probability values are known

In this course, we very briefly talk about the Bayesian decision theory and how to estimate the probabilities from the given data

- CS 551 (Pattern Recognition) course covers these topics thoroughly
- You can also refer to the following books
  - Pattern Classification (by Duda, Hart, and Stork): Chapter 2 for Bayesian decision theory and Chapter 3 for parameter estimation
  - Introduction to Machine Learning (by Alpaydin): Chapter 3 for Bayesian decision theory and Chapter 4 for parametric methods

Bayesian Decision Theory

- Consider a simple decision problem
  Fish classification

- Let's assume that a fish emerges nature in one of the following states
  - **State of nature**

$$C = \begin{cases} C_1 & \text{for } hamsi \\ C_2 & \text{for } barbun \end{cases}$$

- To predict what type will emerge next, we consider C as a random variable, which is described probabilistically
  - **Prior probabilities (a priori probabilities)** $P(C_1)$ and $P(C_2)$ reflect our previous knowledge before the fish appears

$$P(C_1) + P(C_2) = 1 \quad \text{(if no other species exist)}$$

Bayesian Decision Theory

- Let's decide a fish is hamsi or barbun when
  1. We are not allowed to see the fish
  2. We know the prior probabilities
  3. The cost is the same for all incorrect decisions

Decision rule:

$$\text{Select} \quad \begin{cases} hamsi & \text{if} \quad P(C_1) > P(C_2) \\ barbun & \text{otherwise} \end{cases}$$

In this case, we always make the same decision

Bayesian Decision Theory

- We usually have more information for making our decisions
  - For example, we can see the fish and measure its color intensity

- We make this measurement relying on the fact that hamsi and barbun emerge nature in different colors
  - We express this difference in probabilistic terms, considering the color intensity *x* as a continuous random variable, whose distribution depends on the state of nature
  - **Class-conditional probability density functions (likelihoods)** $P(x|C_1)$ and $P(x|C_2)$ give the probability of observing color intensity *x* when the state of nature is C_1 and C_2, respectively

Bayesian Decision Theory

- Now let's combine this measurement with our previous knowledge

Joint probability

$$P(C_j, x) = P(C_j \mid x) \cdot P(x) = P(x \mid C_j) \cdot P(C_j)$$

BAYES FORMULA

$$\underbrace{P(C_j \mid x)}_{\textit{Posterior}} = \frac{\overbrace{P(x \mid C_j)}^{\textit{Likelihood}} \cdot \overbrace{P(C_j)}^{\textit{Prior}}}{\underbrace{P(x)}_{\textit{Evidence}}}$$

$$P(x) = \sum_{j=1}^{N} P(x \mid C_j) \cdot P(C_j)$$

Posterior probabilities (a posteriori probabilities)

$P(C_1|x)$ and $P(C_2|x)$ reflect our beliefs of having a particular fish species when the color intensity of the fish is measured as x

$$\sum_{j=1}^{N} P(C_j \mid x) = 1$$

Bayesian Decision Theory

- Let's decide a fish is hamsi or barbun when
  1. We can see the fish and measure its color x
  2. We know the prior probabilities and likelihoods
  3. The cost is the same for all incorrect decisions

Decision rule:

$$\text{Select} \begin{cases} hamsi & \text{if} \quad P(C_1 \mid x) > P(C_2 \mid x) \\ barbun & \text{otherwise} \end{cases}$$

$$\text{Select} \begin{cases} hamsi & \text{if} \quad P(x \mid C_1) \cdot P(C_1) > P(x \mid C_2) \cdot P(C_2) \\ barbun & \text{otherwise} \end{cases}$$

- *Evidence is unimportant since it is the same for all states of nature*
- *Equal priors → Observing each state of nature is equally likely*
- *Equal likelihoods → Measurement x gives no information*

Bayesian Decision Theory

- We use the Bayes' decision rule as to minimize the probability of error

Decision rule:

$$\text{Select} \begin{cases} hamsi & \text{if} \quad P(C_1 \mid x) > P(C_2 \mid x) \\ barbun & \text{otherwise} \end{cases}$$

$$P(error) = \int_{-\infty}^{\infty} P(error, x)\, dx = \int_{-\infty}^{\infty} P(error \mid x)\, P(x)\, dx$$

For every x, select P(error|x) as small as possible, which corresponds to selecting the state of nature (class) with the highest posterior probability

Bayesian Decision Theory

- Now let's generalize the decision problem

States of nature $\{C_1, C_2, \ldots C_c\}$

Possible actions $\{\alpha_1, \alpha_2, \ldots \alpha_a\}$

Loss function $\lambda(\alpha_i \mid C_j)$

Let $x \in R^d$ be a feature vector in a d - dimensional space

For this x, we would take the action α_i that minimizes

the loss $\lambda(\alpha_i \mid C_j)$ if we knew C_j is its true state of nature

However, we do not know the true state of nature

Thus, we should take the action based on expectations

Bayesian Decision Theory

- The expected loss associated with taking action α_i

$$\underbrace{R(\alpha_i \mid x)}_{\textit{Conditional risk}} = \sum_{j=1}^{C} P(C_j \mid x) \cdot \lambda(\alpha_i \mid C_j)$$

$$P(C_j \mid x) = \frac{P(x \mid C_j) \cdot P(C_j)}{P(x)}$$

- We take the action that minimizes the conditional risk

$$\underbrace{\alpha^*}_{\textit{Optimal action}} = \arg\min_i R(\alpha_i \mid x)$$

The resulting minimum risk R* is called *Bayes risk*

Minimum error-rate classification

- In multi-class classification
  - Each state of nature is usually associated with a class
  - Each action is usually interpreted as deciding on a class

1. Consider the zero-one loss function

$$\lambda(\alpha_i \mid C_j) = \begin{cases} 0 & \text{if } \; i = j \quad \text{(correct classification)} \\ 1 & \text{if } \; i \neq j \quad \text{(all incorrect classifications)} \end{cases}$$

The optimal action is

$$\alpha^* = \arg\max_i \; P(C_i \mid x)$$

Selecting the action that minimizes the conditional risk is equivalent to selecting the action that maximizes the posterior probability

WHY???

Minimum error-rate classification

2. Consider the following loss function

$$\lambda(\alpha_i \mid C_j) = \begin{cases} 0 & \text{if } \; i = j & \text{(correct classification)} \\ \lambda_s & \text{if } \; i \neq j, \; i = 1 \text{ to } C & \text{(all incorrect classifications)} \\ \lambda_r & \text{if } \; i = C+1 & \text{(reject action)} \end{cases}$$

The reject action may be desirable when the misclassification cost is too high

Show that an instance is classified as C_i if only if

1. $P(C_i \mid x) \geq P(C_j \mid x)$

   for all $i \neq j$ and $i \neq C+1$

2. $P(C_i \mid x) \geq 1 - \dfrac{\lambda_r}{\lambda_s}$

What happens when $\lambda_r = 0$

What happens when $\lambda_r > \lambda_s$

Classifiers and discriminant functions

- We may represent a classifier with a set of discriminant functions $g_i(x)$ for i = 1, 2, … C
- We then classify a given instance x with the class C_i for which the discriminant function $g_i(x)$ is the largest
- ***Bayes classifier***
  - Defines a discriminant function using the conditional risk

    $$g_i(x) = -R(\alpha_i \mid x)$$

    $$g_i(x) = P(C_i \mid x) \text{, when 0-1 loss function is used}$$

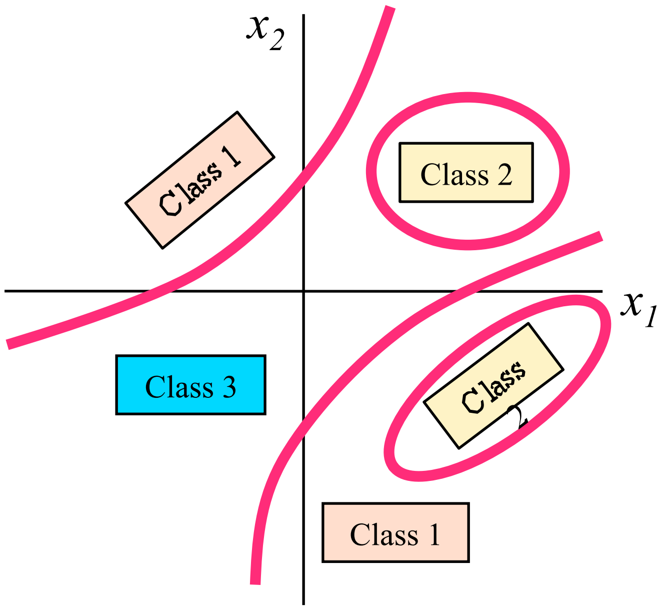
  - Uses the Bayes formula to compute the posteriors

    $$P(C_j \mid x) = \frac{P(x \mid C_j) \cdot P(C_j)}{P(x)}$$

Classifiers and discriminant functions

- We may also define other discriminant functions
  - Linear, quadratic functions
  - Multiplying/shifting the existing ones with positive constants
  - Replacing the existing ones with a monotonically increasing function

$$
\begin{aligned}
g_i(x) &= P(C_i \mid x) \\
&= P(x \mid C_i) \cdot P(C_i) \\
&= \log(P(x \mid C_i) \cdot P(C_i)) \\
&= \underbrace{\log P(x \mid C_i) + \log P(C_i)}_{\textit{Significant simplifications if you use normal distribution}}
\end{aligned}
$$



- Discriminant functions divide the feature space into regions

Classifiers and discriminant functions

- **Discriminant-based approaches** learn discriminant functions directly on the training samples, without estimating class probabilities

- **Likelihood-based approaches** estimate class probabilities on the training samples and then use them to define the discriminant functions

Likelihood-Based Approaches

- **Parametric approach**
  - Assumes a parametric form on the probability distributions and estimate their parameters on the training samples
  - For a given instance x, it estimates the class probabilities of this instance using these distributions
  - Maximum likelihood estimation and Bayesian estimation

- **Nonparametric approach**
  - Does not have such assumption
  - It estimates the class probabilities of the instance x using the nearby points of this instance
  - Parzen windows, k-nearest neighbors

Maximum Likelihood Estimation

- It assumes that the parametric form is known and the parameters are fixed
- It selects the parameters that maximize the likelihood of the training samples

$$\underbrace{P(D \mid \theta)}_{\text{Likelihood of data}} = \prod_{t=1}^{N} P(x^t \mid \theta)$$

Assumes that the training samples are independent and identically distributed

$$\underbrace{\log P(D \mid \theta)}_{\text{Log likelihood of data}} = \sum_{t=1}^{N} \log P(x^t \mid \theta)$$

$$\underbrace{\nabla_{\log P(D|\theta)}}_{\text{Gradient}} = 0$$

How to estimate the parameters of a univariate normal distribution

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right)$$

Naïve Bayes classifier *assumes the independency between every pair of features*