Bayesian Decision Theory

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

Ch. 3 of I2ML

Outline

Introduction

Bayes' Decision Rule

Losses and Risks

Discriminant Functions

Outline

Inference from Data

- Programming computers to make inference from data is a cross between statistics and computer science.
  - statisticians provide the mathematical framework of making inference from data
  - computer scientists work on the efficient implementation of the inference methods
- Data comes from a process that is not completely known.
- This lack of knowledge is indicated by modeling the process as a random process (or stochastic process), which entails probability theory to analyze it.

Coin Tossing Example

- Outcome of tossing a coin $\in$ {head, tail}
- Random variable X:

$$X = \begin{cases} 1 & \text{if outcome is head} \\ 0 & \text{if outcome is tail} \end{cases}$$

- X is Bernoulli-distributed:

$$P(X = x) = p_0^x (1 - p_0)^{1-x}$$

  where the parameter p_0 is the probability that the outcome is head, i.e., $p_0 = P(X = 1)$.

Estimation and Prediction

- Estimation of parameter p_0 from sample $\mathcal{X} = \{x^t\}_{t=1}^N$:

$$\hat{p}_0 = \frac{\#\text{heads}}{\#\text{tosses}} = \frac{\sum_{t=1}^N x^t}{N}$$

- Prediction of outcome of next toss (decision rule):

$$\text{Predicted outcome} = \begin{cases} \text{head} & \text{if } p_0 > 1/2 \\ \text{tail} & \text{otherwise} \end{cases}$$

  by choosing the more probable outcome, which minimizes the probability of error or misclassification error ($= 1 -$ probability of our choice for the predicted outcome).

Credit Scoring Example

- Output: two classes (categories) risk $\in$ {low,high}, or $C \in \{0, 1\}$ (a Bernoulli random variable)
- Prior probabilities: $P(C = 0)$ and $P(C = 1)$ with $P(C = 0) + P(C = 1) = 1$
- Decision rule:

$$\text{Predicted outcome} = \begin{cases} \text{low} & \text{if } P(C = 0) > 1/2 \\ \text{high} & \text{otherwise} \end{cases}$$

  - always predict that the people comes from one class
  - no need to look at the people

Outline

Classification as Bayesian Decision - I

- In the credit scoring example, based on knowledge, we observe customer's yearly income and savings, denoted by two random variables X_1 and X_2.
  - We observe them because we have reason to believe that they give us an idea about the credibility of a customer.
- With the two observables, the credibility of a customer is a Bernoulli random variable C conditioned on the observables $\mathbf{X} = [X_1, X_2]^T$, i.e.,

$$C \mid \mathbf{X}$$

  where $C = 1$ indicates a high-risk customer and $C = 0$ indicates a low-risk customer.
- If we know the distribution $P(C \mid \mathbf{X})$, when a new application arrives with $X_1 = x_1$ and $X_2 = x_2$, we can make a prediction.

Classification as Bayesian Decision - II

- Credit scoring example:
  - Inputs: two features (income and savings), or $\mathbf{x} = [x_1, x_2]^T$
  - Output: two classes (categories) risk $\in \{\text{low,high}\}$, or $C \in \{0, 1\}$ (a Bernoulli random variable)
- Prediction (decision rule):

$$\text{Choose } \begin{cases} C = 1 & \text{if } P(C = 1 \mid \mathbf{x}) > 0.5 \\ C = 0 & \text{otherwise} \end{cases}$$

  or equivalently

$$\text{Choose } \begin{cases} C = 1 & \text{if } P(C = 1 \mid \mathbf{x}) > P(C = 0 \mid \mathbf{x}) \\ C = 0 & \text{otherwise} \end{cases}$$

- Probability of error:

$$1 - \max\big(P(C = 1 \mid \mathbf{x}), P(C = 0 \mid \mathbf{x})\big) = \min\big(P(C = 1 \mid \mathbf{x}), P(C = 0 \mid \mathbf{x})\big)$$

- similar to coin tossing except that C is conditioned on two observable variables $\mathbf{x}$

Bayes' Rule

- Bayes' rule:

$$\text{Posterior } P(C \mid \mathbf{x}) = \frac{\text{likehihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathbf{x} \mid C)P(C)}{p(\mathbf{x})}$$

- prior probability: knowledge we have as to C before looking at the observables $\mathbf{x}$
- posterior probability: knowledge we have as to C after observing $\mathbf{x}$
- class likelihood (class-conditional density): probability of $\mathbf{x}$ given C and derived from data
- evidence: the marginal probability that an observation $\mathbf{x}$ is seen
- Some useful properties to note:
  - $P(C = 0) + P(C = 1) = 1$
  - $p(\mathbf{x}) = p(\mathbf{x} \mid C = 1)P(C = 1) + p(\mathbf{x} \mid C = 0)P(C = 0)$
  - $P(C = 0 \mid \mathbf{x}) + P(C = 1 \mid \mathbf{x}) = 1$
- we will discuss the estimation of $p(C)$ and $p(\mathbf{x}|C)$ from training samples in later lectures

Bayes' Rule for $K > 2$ Classes

- Bayes' rule for general case (K mutually exclusive and exhaustive classes):

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

- Optimal decision rule for Bayes' classifier:

$$\text{Choose } C_i \text{ if } P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$$

  - If the class likelihoods $p(\mathbf{x} \mid C_i)$ are equal, then the decision will rely exclusively on the priors.
  - Conversely, if we have uniform priors $P(C_i)$, then the decision will rely exclusively on the likelihoods.

Outline

Losses and Risks

- In general different decisions or actions may not be equally good or costly.
- Action α_i: decision to assign the input $\mathbf{x}$ to class C_i
- Loss λ_{ik}: loss incurred for taking action α_i when the actual state is C_k
- Expected risk/loss or conditional risk for taking action α_i given input $\mathbf{x}$:

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

- Optimal decision rule with minimum expected risk:

$$\text{Choose } \alpha_i \text{ if } R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$$

0-1 Loss Function

- All correct decisions have zero loss and all errors have unit cost (i.e., are equally costly):

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

- Expected risk:

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

- Optimal decision rule with minimum expected risk (or, equivalently, highest posterior probability):

$$\text{Choose } \alpha_i \text{ if } P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$$

Reject Option - I

- If the certainty of a decision is low but misclassification has very high cost, the action of reject (or doubt), i.e., α_{K+1}, may be more desirable.
- A possible loss function:

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

  where $0 < \lambda < 1$ is the loss incurred for choosing the action of reject.
- Expected risk:

$$R(\alpha_i \mid \mathbf{x}) = \begin{cases} \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda & \text{if } i = K+1 \\ \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x}) & \text{if } i \in \{1, \ldots, K\} \end{cases}$$

Reject Option - II

- Optimal decision rule:

$$\begin{cases} \text{Choose } C_i & \text{if } R(\alpha_i \mid \mathbf{x}) = \min_{1 \le k \le K} R(\alpha_k \mid \mathbf{x}) < R(\alpha_{K+1} \mid \mathbf{x}) \\ \text{Reject} & \text{otherwise} \end{cases}$$

- Equivalent form of optimal decision rule:

$$\begin{cases} \text{Choose } C_i & \text{if } P(C_i \mid \mathbf{x}) = \max_{1 \le k \le K} P(C_k \mid \mathbf{x}) > 1 - \lambda \\ \text{Reject} & \text{otherwise} \end{cases}$$

- This approach is meaningful only if $0 < \lambda < 1$:
  - If $\lambda = 0$, we always reject (a reject is as good as a correct classification).
  - If $\lambda \geq 1$, we never reject (a reject is at least as costly as, or costlier than, a misclassification error).

Outline

Discriminant Functions - I

- One way of specifying a classifier for classification is through a set of discriminant functions,

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

- Classification rule:

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

- Some ways of defining the discriminant functions:
  - $g_i(\mathbf{x}) = -R(\alpha_i \mid \mathbf{x})$ (minimum conditional risk discriminant; generally for Bayes' classifier)
  - $g_i(\mathbf{x}) = P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{\sum_k p(\mathbf{x}|C_k)P(C_k)}$ (minimum error-rate discriminant)

Discriminant Functions - II

- Discriminant functions are not unique.
- Different discriminant functions may correspond to the same decision rule
  - $g_i(\mathbf{x})$ is multiplied by a positive constant
  - $g_i(\mathbf{x})$ is biased by an additive constant
  - $g_i(\mathbf{x})$ is replaced by $f(g_i(\mathbf{x}))$ where $f(\cdot)$ is a monotonically increasing function
- The following all yield the same exact classification results for minimum error-rate classification.
  - $g_i(\mathbf{x}) = P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{\sum_k p(\mathbf{x}|C_k)P(C_k)}$
  - $g_i(\mathbf{x}) = p(\mathbf{x} \mid C_i)P(C_i)$
  - $g_i(\mathbf{x}) = \log p(\mathbf{x} \mid C_i) + \log P(C_i)$

Discriminant Functions - III

- For the two-class case, it suffices to use just one discriminant function:

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

with the following classification rule:

$$\text{Choose } \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

- Various manipulations of the discriminant:
  - $g(\mathbf{x}) = P(C_1 \mid \mathbf{x}) - P(C_2 \mid \mathbf{x})$
  - $g(\mathbf{x}) = \log \frac{p(\mathbf{x}|C_1)}{p(\mathbf{x}|C_2)} + \log \frac{P(C_1)}{P(C_2)}$

Decision Regions

- The feature space is divided into K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$,where

$$\mathcal{R}_i = \left\{ \mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x}) \right\}$$

- The decision region corresponding to a class may consist of noncontiguous subregions.
- The decision regions are separated by decision boundaries (a.k.a. decision surfaces) where ties occur among the discriminant functions with the largest values.