ECE271A: Statistical Learning

Course Webpage

October 15, 2021

Abstract

Warning: This is only a piece of lecture notes written by a careless scribe. So just be careful with and tolerant of any possible typos or misunderstandings when you read [0.1]. The scribe does not intend to make anyone to be driven by his stupidity! Also, the professor's explanation is extremely helpful as he discusses a lot about the interpretable ideas behind the dull scripts. So watch the lecture before reading this. If you have any suggestions (e.g. typos, typography, logistics), please do not hesitate contacting the scribe!

Without specifications, the notation use is as the following

- $\mathbb{R}, \mathbb{C}, \mathbb{Q}, \ldots$: real, complex, quadratic, and so on
- $\mathcal{L}$ – loss; $\mathcal{R}$ – risk

[0.1] Especially '$\cap$' and '$\cup$' are often mistaken because of typos.

Contents

ECE271A **Fall 2021**

Lecture 1: Introduction

Lecturer: Nuno Vasconcelos *Scribes: Rabbittac*

1.1 Introduction

- **Generative**: probability model to fit each class.
- **Discriminant**: decision boundary that separates the classes.

Bayesian Decision Theory is a framework for computing optimal decisions on problems involving uncertainty.

Settings: The world has states or classes drawn from a state or class random variable Y. An observer measures observations drawn from a random process X. The observer uses the observations to make decisions about the state of the world y.

If $x \in \Omega$ and $y \in \Psi$, then the **decision function** is the mapping $g : \Omega \to \Psi$ such that $g(x) = y_o$ and y_o is a prediction of the state y. The **loss function** is the cost $\mathcal{L}(y_o, y)$ of deciding for y_o when the true state is y. And the goal is to determine the optimal decision function for the loss $\mathcal{L}$. In classification, we will mostly consider the **0-1 loss** function

$$\mathcal{L}(g(x), y) = \begin{cases} 1 & g(x) \neq y \\ 0 & g(x) = y \end{cases}$$

In the regression case, the observer tries to predict a continuous y. One of the choice is **square loss**

$$\mathcal{L}(g(x), y) = \|y - g(x)\|^2$$

ECE271A **Fall 2021**

Lecture 2: Bayes Decision Theory

Lecturer: Nuno Vasconcelos *Scribes: Rabbittac*

2.1 Probability

In order to find the optimal decision function, we need a probabilistic description of the problem. Consider the joint distribution $P_{X,Y}(x,i)$. We often decompose it into the combination of two terms

$$P_{X,Y}(x,i) = P_{X,Y}(x|i)P_Y(i)$$

where the first term is referred as **class conditional distribution** and the second term is referred as **class probability**. The class conditional distribution is the model for the observations given the class or state of the world; the class probability is the prior probability of state i and reflects a prior belief that, if all else is equal, the world will be in state i with probability $P_Y(i)$.

Note that by recursive use of the decomposition, we can write

$$\begin{aligned} P_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) =& P_{X_1|X_2,\ldots,X_n}(x_1|x_2,\ldots,x_n) \\ & \times P_{X_2|X_3,\ldots,X_n}(x_2|x_3,\ldots,x_n) \\ & \times \cdots \times P_{X_{n-1}|X_n}(x_{n-1}|x_n)P_{X_n}(x_n) \end{aligned}$$

e.g.1. In the medical diagnosis scenario, the probability that you will be sick and have 104deg of fever will be $P_{Y,X_1}(\text{sick},104) = P_{Y|X_1}(\text{sick}|104)P_{X_1}(104)$. The chain rule breaks the original problem into two easier ones: $P(\text{sick}|104) \approx 1$ and $P(104)$ can be computed by sampling and making a histogram.

In many occasions, we have problems with multiple random variables. We can summarize this as a vector $X = (x_1,\ldots,x_n)$ of n random variables. But often we care about only a subset of these random variable. This is done by **marginalization**

$$P_{X_i,X_j}(x_i,x_j) = \sum_{X \setminus \{x_i,x_j\}} P_{X_1,\ldots,X_n}(x_1,\ldots,x_n)$$

Typically, we combine with the chain rule to explore independence relationships that will allow us to reduce computation, where X and Y are **independent** if

$$P_{X,Y}(x,y) = P_X(x)P_Y(y)$$

e.g.2. To compute $P_Y(\text{sick})$, (1) by marginalization, $P_Y(\text{sick}) = \sum_s \int P_{Y,X_1,S}(\text{sick}, x, s)\mathrm{d}x = \sum_s \int P_{Y|X_1,S}(\text{sick}|x,s)P_{S|X_1}(s|x)P_{X_1}(x)\mathrm{d}x = \int P_{Y|X_1}(\text{sick}|x)P_{X_1}(x) \sum_s P_{S|X_1}(s|x)\mathrm{d}x = \int P_{Y|X_1}(\text{sick}|x)P_{X_1}(x)\mathrm{d}x.$

The central equation of Bayesian inference is the **Bayes Rule**

$$P_{Y|X}(y|x) = \frac{P_{X|Y}(x|y)P_Y(y)}{P_X(x)}$$

which allows us to switch the relationship between the variables.

e.g.3. For medical diagnosis doctor, he needs to know $P_{Y|X}(\text{disease } y|\text{symptom } x)$. By Bayes Rule, $P_{Y|X}(\text{disease } y|\text{symptom } x) = \frac{P_{X|Y}(\text{symptom } x|\text{disease } y)P_Y(\text{disease } y)}{P_X(\text{symptom } x)}$, where $P_{X|Y}(\text{symptom } x|\text{disease } y)$ can be got from medical textbook, P_Y(disease y does not depend on the patient and $P_X(\text{symptom } x)$ is a combination of the two marginalization.

2.2 Bayes Decision Rule

The **risk** is defined as the expected value of the loss

$$\mathcal{R} = \mathbb{E}_{X,Y}[\mathcal{L}(X,Y)] = \int \sum_{i=1}^{M} P_{Y,X}(i,x)\mathcal{L}(g(x),i)\mathrm{d}x$$

where X denotes observations, Y denotes the state of the world, and $g(\cdot)$ is our decision function.

Now by chain rule,

$$\begin{aligned}
\mathcal{R} &= \int P_X(x) \sum_{i=1}^{M} P_{Y|X}(i|x)\mathcal{L}(g(x),i)\mathrm{d}x \\
&= \int P_X(x)R(x)\mathrm{d}x \\
&= \mathbb{E}_X[R(x)]
\end{aligned}$$

where

$$R(x) = \sum_{i}^{M} P_{Y|X}(i|x)\mathcal{L}(g(x),i)$$

is the **conditional risk** given the observation x.

By definition, $\forall x,y : \mathcal{L}[g(x),y] \geq 0$. Hence, $\forall x : R(x) = \sum_{i=1}^{M} P_{Y|X}(i|x)\mathcal{L}(g(x),i) \geq 0$. Also, $\mathcal{R} = \mathbb{E}_X[R(X)]$ is minimum if we minimize $R(x)$ at all x. i.e. the decision function is

$$g^*(x) = \arg\min_{g(x)} \sum_{i=1}^{M} P_{Y|X}(i|x)\mathcal{L}(g(x),i)$$

which is the **Bayes Decision Rule**. The associate risk is called **Bayes risk**

$$\begin{aligned}
R^* &= \int \sum_{i=1}^{M} P_{Y,X}(i,x)\mathcal{L}(g^*(x),i)\mathrm{d}x \\
&= \int P_X(x) \sum_{i=1}^{M} P_{Y|X}(i|x)\mathcal{L}(g^*(x),i)\mathrm{d}x
\end{aligned}$$

2.2.1 BDR for Binary Classification

For a binary classification problem $g^*(x) \in \{0,1\}$, the conditional risk is $R(x) = \sum_{i=0}^{1} P_{Y|X}(i|x)\mathcal{L}(g(x),i) = P_{Y|X}(0|x)\mathcal{L}(g(x),0) + P_{Y|X}(1|x)\mathcal{L}(g(x),1)$. We have two options: $g(x) = 0 \implies R_0(x) = P_{Y|X}(0|x)\mathcal{L}(0,1) + P_{Y|X}(1|x)\mathcal{L}(0,1)$ and $g(x) = 1 \implies R_1(x) = P_{Y|X}(0|x)\mathcal{L}(0,1) + P_{Y|X}(1|x)\mathcal{L}(0,1)$. Then pick the one with smaller conditional risk. i.e. Pick

$$\begin{cases} g(x) = 0 & R_0(x) < R_1(x) \\ g(x) = 1 & R_0(x) > R_1(x) \end{cases}$$

Or we can write: pick 0 if

$$\frac{P_{Y|X}(0|x)}{P_{Y|X}(1|x)} > \frac{\mathcal{L}(0,1)}{\mathcal{L}(1,0)}$$

Apply Bayes rule, $\frac{P_{X|Y}(x|0)P_Y(0)}{P_{X|Y}(x|1)P_Y(1)} > \frac{\mathcal{L}(0,1)}{\mathcal{L}(1,0)}$, which is equivalent to pick 0 if

$$\frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)} > T^* = \frac{\mathcal{L}(0,1)P_Y(1)}{\mathcal{L}(1,0)P_Y(0)}$$

2.3 BDR for 0-1 Loss & MAP

Let's consider the 0-1 loss $\mathcal{L}(g(x),y) = \begin{cases} 1 & g(x) \neq y \\ 0 & g(x) = y \end{cases}$. In this case, the optimal decision function is

$$\begin{aligned} g^*(x) &= \arg\min_{g(x)} \sum_{i=1}^{M} P_{Y|X}(i|x)\mathcal{L}(g(x),i) \\ &= \arg\min_{g(x)} \sum_{i \neq g(x)} P_{Y|X}(i|x) \\ &= \arg\min_{g(x)} 1 - P_{Y|X}(g(x)|x) \\ &= \arg\max_{g(x)} P_{Y|X}(g(x)|x) \\ &= \arg\max_{i} P_{Y|X}(i|x) \end{aligned}$$

The associated risk is

$$R^* = \int P_{Y,X}(y \neq g^*(x), x)\mathrm{d}x$$

For 0-1 loss, the Bayes decision rule is called **maximum a posterior (MAP)**. The risk is the probability of error of this rule. Apply Bayes rule, MAP can be generalized to

$$i^*(x) = \arg\max_{i} P_{X|Y}(x|i)P_Y(i)$$

Note that in machine learning, many class-conditional distributions are exponential (e.g. the Gaussian), where the product can be tricky to compute (e.g. the tail probabilities are quite small). We can take advantage of the fact that we only care about the order of the terms on the right-hand side. So take log, then

$$\begin{aligned} i^*(x) &= \arg\max_i P_{X|Y}(x|i)P_Y(i) \\ &= \arg\max_i \log\big(P_{X|Y}(x|i)P_Y(i)\big) \\ &= \arg\max_i \log\big(P_{X|Y}(x|i) + \log(P_Y(i))\big) \end{aligned}$$

These three rules are equivalent but the first one is often hard to use, and the last one is more frequent.

ECE271A **Fall 2021**

Lecture 3: Bayesian Decision Theory & Gaussian Classifier

Lecturer: Nuno Vasconcelos *Scribes: Rabbittac*

3.1 BDT for Gaussian

e.g.1. In communications, a bit is transmitted by a source, corrupted by noise, and received by a decoder. Q: what should the optimal decoder do to recover Y?

This was modeled as a classification problem with Gaussian classes where $P_{X|Y}(x|0) = \mathcal{G}(x, \mu_0, \sigma)$ and $P_{X|Y}(x|1) = \mathcal{G}(x, \mu_1, \sigma)$ where $P_Y(0) = P_X(1) = \frac{1}{2}$ for which the optimal decision boundary is a threshold $x < \frac{\mu_1+\mu_0}{2}$ [3.1].

If the class probabilities are not the same such that $P_X(1) > P_Y(0)$, then the decision rule will be

$$\begin{aligned}
i^*(x) &= \arg\max_i \big((\log P_{X|Y}(x|i) + \log P_Y(i)\big) \\
&= \arg\max_i \left(\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}\right) + \log P_Y(i)\right) \\
&= \arg\min_i \left(\frac{(x-\mu_i)^2}{2\sigma^2} - \log P_Y(i)\right) \\
&= \arg\min_i \left(-2x\mu_i + \mu_i^2 - 2\sigma^2 \log P_Y(i)\right)
\end{aligned}$$

So the optimal decision is to pick 0 if

$$x < \frac{\mu_1 + \mu_0}{2} + \frac{\sigma^2}{\mu_1 - \mu_0} \log \frac{P_Y(0)}{P_Y(1)}$$

where the term $\frac{\mu_0+\mu_1}{2}$ is the observations and the product is of problem difficulty and prior knowledge [3.2]. Note that if $P_Y(0) = 1$, then we only care about $Y = 0$ so the threshold is infinite.

3.2 Multivariate Gaussian Classifier

In practice, we rarely have single variables. Typically, data $X = (X_1, X_n)$ is a vector of observations. BDR for this case is equivalent except that the class-conditional distribution becomes **multivariate Gaussian**

$$P_{X|Y}(x|i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\right)$$

[3.1] In this example, we made assumptions that (1) uniform class probabilities, (2) Gaussian, (3) the same variance of two distributions (4) additive noise.

[3.2] It is weighed by the inverse of the normalized distance between the means.

In this case, the BDR is

$$i^*(x) = \arg\max_i \left(-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log(2\pi)^d|\Sigma_i| + \log P_Y(i) \right)$$

3.2.1 BDR for n-d Gaussian with Equal Covariance

We first consider a special case of interest that all classes have the same covariance i.e. $\forall i : \Sigma = \Sigma_i$. In this case, BDR can be written as

$$i^*(x) = \arg\min_i \left(d_i(x, \mu_i) + \alpha_i \right)$$

where $d(x,y) = (x-y)^T\Sigma^{-1}(x-y)$ is Mahalanobis distance and $\alpha_i = -2\log P_Y(i)$ is added to account for the class prior. In detail,

$$\begin{aligned}
i^*(x) &= \arg\min_i \left((x-\mu_i)^T\Sigma^{-1}(x-\mu_i) - 2\log P_Y(i) \right) \\
&= \arg\min_i \left(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_i - \mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i - 2\log P_Y(i) \right) \\
&= \arg\min_i \left(x^T\Sigma^{-1}x - 2\mu_i{}^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i - 2\log P_Y(i) \right) \\
&= \arg\max_i \left(\mu_i^T\Sigma^{-1}x - \frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \log P_Y(i) \right)
\end{aligned}$$

In summary, when classes have equal covariance, BDR is a linear function or a **linear discriminant** $i^*(x) = \arg\max_i g_i(x)$ where $g_i(x) = w_i^T x + w_{i0}$ for $w_i^T = \mu_i^T\Sigma^{-1}x$ and $w_{i0} = -\frac{1}{2}\mu_i^T\Sigma^{-1}\mu_i + \log P_Y(i)$. Applying $\mu_i^T\Sigma^{-1}\mu_i - \mu_j^T\Sigma^{-1}\mu_j = (\mu_i+\mu_j)^T\Sigma^{-1}(\mu_i-\mu_j)$, we can rewrite this as

$$(\mu_i - \mu_j)^T\,\Sigma^{-1}x - \frac{1}{2}\left(\mu_i^T\Sigma^{-1}\mu_i - \mu_j^T\Sigma^{-1}\mu_j - 2\log\frac{P_Y(i)}{P_Y(j)} \right) = 0$$

The geometric interpretation is that the linear discriminant can be written as a hyperplane of normal vector w that passes through x_0

$$w^T(x - x_0) = 0$$

with

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\mu_i - \mu_j}{(\mu_i-\mu_j)^T\Sigma^{-1}(\mu_i-\mu_j)}\log\frac{P_Y(i)}{P_Y(j)}$$

3.2.2 BDR for n-d Gaussian with $\Sigma = \sigma^2 I$

Another special case is when $\Sigma = \sigma^2 I$. Then the optimal boundary has

$$w = \frac{\mu_i - \mu_j}{\sigma^2}$$

$$x_0 = \frac{\mu_i + \mu_j}{2} - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2}\log\frac{P_Y(i)}{P_Y(j)}(\mu_i - \mu_j)$$

where w has the geometric interpretation that it is the vector along the line through μ_i and μ_j