CS5487 Problem Set 6

Bayes Decision Theory

Antoni Chan
Department of Computer Science
City University of Hong Kong

Bayes Decision Theory

Problem 6.1 BDR with unbalanced loss function

Consider a two-class problem with $y \in \{0, 1\}$ and measurement x, with associated prior distribution $p(y)$ and class-conditional densities $p(x|y)$.

(a) Consider the loss-function:

$$L(g(x), y) = \begin{cases} 0, & g(x) = y \\ \ell_0, & y = 0 \text{ and } g(x) = 1 \\ \ell_1, & y = 1 \text{ and } g(x) = 0. \end{cases} \tag{6.1}$$

In other words, the loss for misclassification is different for each class. When might this type of loss function be useful? Can you give a real-world example?

(b) Derive the Bayes decision rule (BDR) for y. Write the BDR as a log-likelihood ratio test. What is the threshold?

(c) Explain how the loss values ℓ_0 and ℓ_1 influence the threshold.

.....

Problem 6.2 BDR for regression

In this problem, we will consider the Bayes decision rule for regression. Suppose we have a regression problem, where $y \in \mathbb{R}$ is the output, $x \in \mathbb{R}^d$ is the input, and we have already learned the distribution $p(y|x)$, which maps the input x to a distribution of outputs y. The goal is to select the optimal output y for a given x.

(a) Consider the squared-loss function, $L(g(x), y) = (g(x) - y)^2$. Show that the BDR is to decide the conditional mean of $p(y|x)$, or $g^*(x) = \mathbb{E}[y|x]$. In other words, show that $g^*(x)$ minimizes the conditional risk $R(x) = \int L(g(x), y)p(y|x)dy$.

(b) One generalization of the squared-loss function is the *Minkowski* loss,

$$L_q(g(x), y) = |g(x) - y|^q. \tag{6.2}$$

Plot the loss function L_q versus $(g(x) - y)$ for values of $q \in \{0.2, 1, 2, 10\}$ and $q \to 0$. Comment on the effect of using different loss functions.

(c) Show that the BDR for $q = 1$ is to select the conditional median of $p(y|x)$.

(d) Show that the BDR for $q \to 0$ is to select the conditional mode of $p(y|x)$.

.....

Problem 6.3 Noisy channel with unequal priors

In this problem, you will derive the BDR for a noisy channel with unequal priors (an example in lecture). Given a bit $y = \{0, 1\}$, the transmitter sends a signal of μ_0 for $y = 0$, and μ_1 for $y = 1$. The channel has additive zero-mean Gaussian noise with variance σ^2, and hence the class-conditional densities for the measurement $x \in \mathbb{R}$ are

$$p(x|y = 0) = \mathcal{N}(x|\mu_0, \sigma^2), \quad p(x|y = 1) = \mathcal{N}(x|\mu_1, \sigma^2). \tag{6.3}$$

Let the prior probability of the transmitted bits be $p(y = 1) = \pi_1$ and $p(y = 0) = \pi_0$. The goal is to recover the bit $y \in \{0, 1\}$ after receiving a noisy measurement x.

(a) Show that the Bayes decision rule (BDR) using the 0-1 loss function is given by

$$y^* = \begin{cases} 0, & x < \frac{\mu_1+\mu_0}{2} + \frac{\sigma^2}{\mu_1-\mu_0} \log \frac{\pi_0}{\pi_1} \\ 1, & \text{otherwise} \end{cases} \tag{6.4}$$

(b) Explain the intuitive effect of each term in the above BDR.

(c) Let $\mu_0 = -1$, $\mu_1 = 1$, and $\sigma^2 = 1$. Plot the change in the threshold for different values of π.

.....

Problem 6.4 Coin Tossing

In this problem we will consider the traditional probability scenario of coin tossing. However, we will consider two variations. First, the coin is not fair. Denoting by s the outcome of the coin toss (either H or T) we have

$$p(s = H) = \alpha, \; \alpha \in [0, 1]. \tag{6.5}$$

Second, you do not observe the coin directly but have to rely on a friend that reports the outcome of the toss. Unfortunately your friend is unreliable, he will sometimes report heads when the outcome was tails and vice-versa. Denoting the report by r we have

$$p(r = T|s = H) = \theta_1, \tag{6.6}$$
$$p(r = H|s = T) = \theta_2, \tag{6.7}$$

where $\theta_1, \theta_2 \in [0, 1]$. Your job is to, given the report from your friend, guess the outcome of the toss.

(a) Given that your friend reports heads, what is the optimal decision function in the minimum probability of error sense? That is, when should you guess heads, and when should you guess tails?

(b) Consider the case $\theta_1 = \theta_2$. Can you give an intuitive interpretation to the rule derived in (a)?

(c) You figured out that if you ask your friend to report the outcome of the toss various times, he will produce reports that are statistically independent. You then decide to ask him to report the outcome n times, in the hope that this will reduce the uncertainty. (Note: there is still only one coin toss, but the outcome gets reported n times). What is the new minimum probability of error decision rule?

(d) Consider the case $\theta_1 = \theta_2$ and assume that the report sequence is *all heads*. Can you give an intuitive interpretation to the rule derived in (c)?

.....

Problem 6.5 Naive Bayes and discrete variables

For high-dimensional observation spaces, it might be difficult to learn a joint density over the space (e.g., if not enough data is available). One common assumption is to use a "Naive Bayes" model, where we assume that the individual features (dimensions) are conditionally independent given the class,

$$p(x|y=j) = \prod_{i=1}^{n} p(x_i|y=j), \tag{6.8}$$

where $x = [x_1, \cdots, x_d]^T$ is the observation vector, and x_i is the individual feature. While the features are conditionally independent given the class, the features are still dependent in the overall distribution of observations $p(x)$ (similar to a GMM with diagonal covariance matrices).

Let the vector x be a collection of d binary-valued features, i.e.

$$x = [x_1, \cdots, x_d]^T, \quad x_i \in \{0, 1\}. \tag{6.9}$$

Assume there are C classes, with class variable $y \in \{1, \cdots, C\}$ and prior distribution $p(y=j) = \pi_j$. Now define

$$p_{ij} = p(x_i = 1|y=j), \ \forall i, j \tag{6.10}$$

with the features $\{x_i\}$ being conditionally independent given class $y = j$ (this conditional independence assumption is the "Naive Bayes" assumption). The goal is to recover the class y given a measurement x.

(a) Interpret in words the meaning of p_{ij}.

(b) Show that the class-conditional distributions can be written as

$$p(x|y=j) = \prod_{i=1}^{d} p_{ij}^{x_i}(1-p_{ij})^{1-x_i}. \tag{6.11}$$

(c) Show that the minimum probability of error is achieved by the following decision rule: Decide $y = j$ if $g_j(x) \geq g_k(x)$ for all j, k, where

$$g_j(x) = \sum_{i=1}^{d} x_i \log \frac{p_{ij}}{1-p_{ij}} + \sum_{i=1}^{d} \log(1-p_{ij}) + \log \pi_j. \tag{6.12}$$

(d) Now assume there are two classes, $C = 2$, and for convenience let $p_i = p_{i1}$ and $q_i = p_{i2}$. Using the above result, show that the BDR is: Decide $y = 1$ if $g(x) > 0$, and $y = 2$ otherwise, where

$$g(x) = \sum_{i=1}^{d} \left[x_i \log \frac{p_i}{q_i} + (1 - x_i) \log \frac{1 - p_i}{1 - q_i} \right] + \log \frac{\pi_1}{\pi_2}. \tag{6.13}$$

(e) Show that $g(x)$ can be rewritten as

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0, \tag{6.14}$$

where

$$w_i = \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)}, \qquad w_0 = \sum_{i=1}^{d} \log \frac{1 - p_i}{1 - q_i} + \log \frac{\pi_1}{\pi_2}. \tag{6.15}$$

This demonstrates that the BDR in this case is a simple linear classifier, where each feature "votes" for the class using its weight w_i. What is the interpretation of formulas for the weights w_i and the offset w_0?

.....

Gaussian Classifiers

Problem 6.6 Gaussian classifier with common covariance

In this problem, we will derive the BDR for Gaussian classifiers with a common covariance, and interpret the resulting decision boundaries. Let $y \in \{1, \ldots, C\}$ be the classes with prior probabilities $p(y = j) = \pi_j$, and $x \in \mathbb{R}^d$ be the measurement with class conditional densities that are Gaussian with a shared covariance, $p(x|y = j) = \mathcal{N}(x|\mu_j, \Sigma)$.

(a) Show that the BDR using the 0-1 loss function is:

$$g(x)^* = \operatorname*{argmax}_j g_j(x), \tag{6.16}$$

where the $g_j(x)$ for each class is a linear function of x,

$$g_j(x) = w_j^T x + b_j, \tag{6.17}$$

$$w_j = \Sigma^{-1}\mu_j, \quad b_j = -\frac{1}{2}\mu_j^T \Sigma^{-1}\mu_j + \log \pi_j. \tag{6.18}$$

(b) For two classes i and j ($i \neq j$), show that the decision boundary between the two classes (i.e., $g_i(x) = g_j(x)$) is described by a hyperplane,

$$w^T x + b = 0, \tag{6.19}$$

$$w = \Sigma^{-1}(\mu_i - \mu_j), \quad b = -\frac{1}{2}(\mu_i + \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j) + \log \frac{\pi_i}{\pi_j} \tag{6.20}$$

(c) Finally, show that the hyperplane in (6.19) can be rewritten in the form

$$w^T(x - x_0) = 0, \tag{6.21}$$

$$w = \Sigma^{-1}(\mu_i - \mu_j), \quad x_0 = \frac{\mu_i + \mu_j}{2} - \frac{(\mu_i - \mu_j)}{\|\mu_i - \mu_j\|_\Sigma^2} \log \frac{\pi_i}{\pi_j}. \tag{6.22}$$

What is the interpretation of w and x_0? What is the effect of the priors $\{\pi_i, \pi_j\}$ on x_0?

.....

Problem 6.7 Gaussian classifier with arbitrary covariances

In this problem, we will derive the BDR for Gaussian classifiers with *arbitrary* covariances, and interpret the resulting decision boundaries. Consider the same setup as Problem 6.6, but with class conditional densities with individual covariance matrices, $p(x|y = j) = \mathcal{N}(x|\mu_j, \Sigma_j)$.

(a) Show that the BDR using the 0-1 loss function is:

$$g(x)^* = \operatorname*{argmax}_j g_j(x), \tag{6.23}$$

where the $g_j(x)$ for each class is a quadratic function of x,

$$g_j(x) = x^T A_j x + w_j^T x + b_j, \tag{6.24}$$

$$A_j = -\frac{1}{2}\Sigma_j^{-1}, \quad w_j = \Sigma_j^{-1}\mu_j, \quad b_j = -\frac{1}{2}\mu_j^T\Sigma_j^{-1}\mu_j - \frac{1}{2}\log|\Sigma_j| + \log \pi_j. \tag{6.25}$$

(b) For two classes i and j ($i \neq j$), the decision boundary between the two classes, i.e., $g_i(x) = g_j(x)$, is either a hyperplane, sphere, ellipsoid, 2 hyperplanes, parabola, or hyperbola (i.e., a conic section). What parameters of the Gaussian class conditional densities $\{\mu_i, \Sigma_i, \mu_j, \Sigma_j\}$ will give each form?

.....

Problem 6.8 Posterior distribution of binary Gaussian classifier

In the previous 2 problems, we have looked at the log-based version of the BDR with the 0-1 loss function. It is also interesting to look at the original definition of the BDR that is based on posterior distributions,

$$g(x)^* = \operatorname*{argmax}_j p(y = j|x). \tag{6.26}$$

Consider a two-class problem, $y \in \{0, 1\}$ with priors $p(y = 1) = \pi_1$ and $p(y = 0) = \pi_0$. The observation is $x \in \mathbb{R}$, with Gaussian class-conditional densities with equal variance, $p(x|y) = \mathcal{N}(x|\mu_j, \sigma^2)$.

(a) Show that the posterior distribution $p(y|x)$ can be written as

$$p(y = 1|x) = \frac{1}{1 + e^{-f(x)}}, \qquad p(y = 0|x) = \frac{1}{1 + e^{f(x)}}, \tag{6.27}$$

where

$$f(x) = \log p(x|y = 1) - \log p(x|y = 0) + \log \frac{\pi_1}{\pi_0}. \tag{6.28}$$

(b) For the Gaussian CCDs assumed above, show that

$$f(x) = \frac{1}{\sigma^2}\left[(\mu_1 - \mu_0)x - \frac{1}{2}(\mu_1^2 - \mu_0^2)\right] + \log\frac{\pi_1}{\pi_0}. \tag{6.29}$$

Hence, the posterior probability of class 1 given x is a *sigmoid function* of the form $\frac{1}{1+e^{-(ax+b)}}$.

(c) What is the decision boundary for this classifier?

(d) Plot $p(y = 1|x)$ when $\mu_0 = -1$, $\mu_1 = 1$, and $\sigma^2 = 1$, for different values of π_1. Also plot the class-conditional densities.

.....

Error Bounds

Problem 6.9 Error bounds for classification with Gaussians

In this problem, we will derive error bounds for classification using Gaussian class-conditionals. Consider a two-class problem, with $y \in \{0, 1\}$, priors $p(y = 1) = \pi$ and $p(y = 0) = 1 - \pi$, and conditional densities $p(x|y) = \mathcal{N}(x|\mu_y, \Sigma_y)$.

For the 0-1 loss function, the BDR is to select the class with highest posterior probability (or minimum probability of error),

$$y^* = \underset{j}{\operatorname{argmin}}[1 - p(y \neq j|x)] = \underset{j}{\operatorname{argmax}}\, p(y = j|x). \tag{6.30}$$

Hence, for a given x, the probability of error using BDR is

$$p(error|x) = \min[p(0|x), p(1|x)]. \tag{6.31}$$

The total probability of error is then obtained by taking the expectation over x,

$$p(error) = \int p(x)p(error|x)dx. \tag{6.32}$$

(This is also the Bayes risk). Because of the discontinuous nature of the decision regions, the integral in (6.32) is almost always difficult to calculate. For a few dimensions, numerical integration methods could be used, but these become less viable for high dimensional spaces.

In this problem, we will derive a bound on the probability of error, given the above assumptions.

(a) First, verify that the following inequality is true:

$$\min[a, b] \leq a^\beta b^{1-\beta}, \text{ for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1. \tag{6.33}$$

Hint: without loss of generality, consider the case $a < b$.

(b) Apply the bound in (6.33) to (6.32) to show that

$$p(error) \leq (1 - \pi)^\beta \pi^{(1-\beta)} e^{-k(\beta)}, \tag{6.34}$$

where

$$k(\beta) = -\log \int p(x|0)^{\beta} p(x|1)^{1-\beta} dx. \tag{6.35}$$

This is called the *Chernoff bound*. After finding a closed-form expression for $k(\beta)$, we could plugin for our parameters and find a value of β that minimizes the upper bound $e^{-k(\beta)}$, as illustrated in the below figure. (You don't need to do this.)

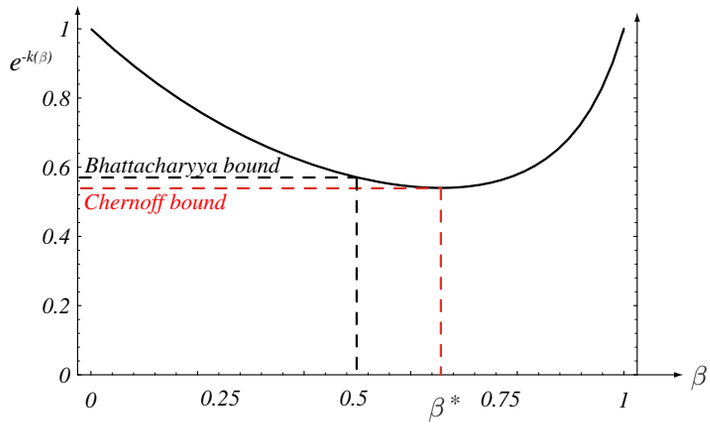

FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Alternatively, we could simply fix β, which yields a looser bound. For $\beta = 1/2$, this is called the *Bhatacharyya* bound.

(c) Consider the case when $\beta = 1/2$, which is the Bhatacharyya bound,

$$p(error) \leq \sqrt{\pi(1-\pi)} e^{-k(\frac{1}{2})}. \tag{6.36}$$

Show that the exponent term is

$$k(\tfrac{1}{2}) = \frac{1}{8}(\mu_1 - \mu_0)^T \left[\frac{\Sigma_1 + \Sigma_0}{2}\right]^{-1} (\mu_1 - \mu_0) + \frac{1}{2} \log \frac{\left|\frac{\Sigma_1+\Sigma_0}{2}\right|}{\sqrt{|\Sigma_1|\,|\Sigma_2|}}. \tag{6.37}$$

Hint: note that $p(x|0)^{\frac{1}{2}}$ can be rewritten as a scaled Gaussian, and apply Problem 1.9.

(d) Describe an intuitive explanation for each of the terms in (6.36) and (6.37) . When is the probability of error large? When is it low? What happens when $\Sigma_1 = \Sigma_0$?

(e) Consider the noisy channel in Problem 6.3. Compute the Bhatacharyya bound when $\pi = 0.5$. Compare the bound to an estimate of the probability of error using numerical integration.

. . . . . .