EECS 553: Machine Learning (ECE) **University of Michigan**

Bayes Classifiers

Winter 2023 **Clayton Scott**

1 Probabilistic Setting for Classification

Probability theory provides a rigorous setting for machine learning. The reader is assumed to be familiar with the following concepts from probability: Random variables (discrete and continuous), jointly distributed random variables, (joint) pdf/ pmf, multivariate Gaussian distribution, independence, conditional distributions, Bayes rule, and the laws of total probability and expectation.

Let $\boldsymbol{X} = [X_1, \ldots, X_d]^T \in \mathbb{R}^d$ be a *feature vector* for a classification problem, and let $Y \in \{1, \ldots, K\}$ be the associated class label. For example, each X_j may be a different attribute of an individual applying to receive a loan, and $Y \in \{1, 2\}$ where $Y = 1$ if the individual will make all of their payments on time, and $Y = 2$ otherwise. We use capital letters for the feature vector and label to emphasize that these quantities are viewed as random variables.

In these notes, and throughout much of this course, we make the following assumption:

> $\boldsymbol{X}$ and Y are *jointly distributed* random variables.

There are two ways in which we will think about the joint distribution. The first is in terms of the marginal distribution of Y, together with the conditional distribution of $\boldsymbol{X}|Y$. The second is in terms of the marginal distribution of $\boldsymbol{X}$, together with the conditional distribution of $Y|\boldsymbol{X}$. These two perspectives are equivalent but one is typically easier to work with depending on the setting. We will use both perspectives.

Example 1. Figure 1 show a joint distribution for a binary ($K = 2$) classification problem with a $d = 2$ dimensional feature space. In this case, the marginal distribution of Y is uniform on the values 1 and 2. The conditional distribution of $\boldsymbol{X}|Y$ is

$$\boldsymbol{X}|Y = 1 \sim N(\boldsymbol{\mu}_1, \Sigma)$$
$$\boldsymbol{X}|Y = 2 \sim N(\boldsymbol{\mu}_2, \Sigma)$$

where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \text{ and } \Sigma = \begin{bmatrix} .9 & .4 \\ .4 & .3 \end{bmatrix}.$$

The marginal distribution of Y and the conditional distribution of $\boldsymbol{X}|Y$ provide a complete characterization of the joint distribution. These quantities allow us to answer any probabilistic question about the data. For example, suppose we wish to know the probability that $X_2 \leq 0.5$. Then

$$\begin{aligned} \Pr(X_2 \leq 0.5) &= \Pr(X_2 \leq 0.5|Y = 1) \Pr(Y = 1) + \Pr(X_2 \leq 0.5|Y = 2) \Pr(Y = 2) \\ &= 0.5\Phi(0.5; 1, \sqrt{.3}) + 0.5\Phi(0.5; -1, \sqrt{.3}) \\ &= 0.5888 \end{aligned}$$

where $\Phi(x; \mu, \sigma)$ is the CDF of a $\mathcal{N}(\mu, \sigma^2)$ random variable.

It is also possible to represent this joint distribution using the marginal distribution of $\boldsymbol{X}$ and the conditional distribution of $Y|\boldsymbol{X}$. By the law of total probability, the marginal density of $\boldsymbol{X}$ is

$$\frac{1}{2}\phi(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}\phi(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma)$$

where $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ is the pdf of a $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ random variable. By Bayes rule, the conditional distribution of $Y|\boldsymbol{X} = \boldsymbol{x}$ is

$$\Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\frac{1}{2}\phi(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma)}{\frac{1}{2}\phi(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}\phi(x; \boldsymbol{\mu}_2, \Sigma)}, \qquad \Pr(Y = 2|\boldsymbol{X} = \boldsymbol{x}) = \frac{\frac{1}{2}\phi(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma)}{\frac{1}{2}\phi(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma) + \frac{1}{2}\phi(x; \boldsymbol{\mu}_2, \Sigma)}.$$

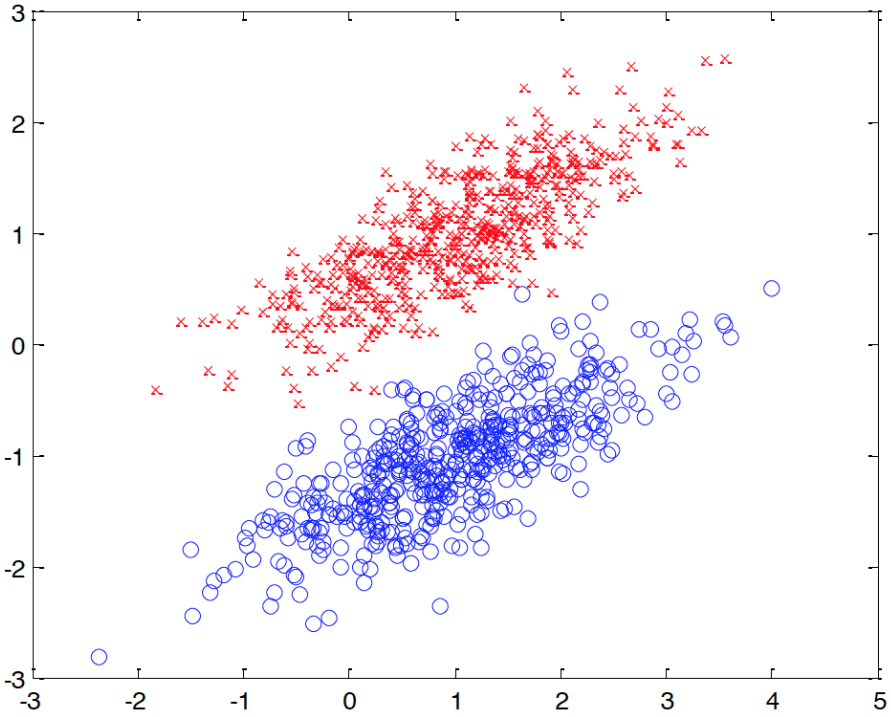

Figure 1: Two-dimensional data with Gaussian class-conditional distributions.

2 The Bayes Classifier

Given a joint distribution $(\boldsymbol{X}, Y)$, we may ask "what is the best possible classifier for that joint distribution?" To answer this question, we need to define what is meant by "best," which means we need to define a performance measure.

A classifier is a function $f : \mathbb{R}^d \to \{1, \ldots, K\}$. The most common performance measure in classification is the *probability of error*, defined by:

$$R(f) := \Pr(f(\boldsymbol{X}) \neq Y).$$

In words, $R(f)$ is the probability that if we draw a random $(\boldsymbol{X}, Y)$ from the joint distribution, that f predicts the wrong label for $\boldsymbol{X}$.

The probability of error is an example of a *risk*, a concept that will be defined later in the course. For now, we use the term "risk" to refer to the probability of error. The *Bayes risk* is the smallest risk of any classifier, and is denoted

$$R^* := \min_{\text{all } f} R(f),$$

where the minimization is over all possible classifiers. If $R(f) = R^*$, f is called a *Bayes classifier*. It is important to note that R^* may not be zero. There are classification problems where the best classifier still makes errors. This occurs when there are feature vectors that could have arisen from multiple different classes.

We introduce the following terminology. The marginal distribution of Y is called the *prior distribution*, while the conditional distribution of $Y|\boldsymbol{X} = \boldsymbol{x}$ is called the *posterior distribution*. The conditional distributions of $\boldsymbol{X}|Y$ are called *class-conditional distributions*.

Let $\pi_k = \Pr(Y = k)$ denote the prior class probabilities, $g_k(\boldsymbol{x})$ the class-conditional pmfs/pdfs of $\boldsymbol{X} \mid Y = k$, and

$$\eta_k(\boldsymbol{x}) = \Pr(Y = k \mid \boldsymbol{X} = \boldsymbol{x})$$

the posterior class probabilities. Notice that $\sum_i \pi_k = 1$ and that $\forall x, \sum_k \eta_k(\boldsymbol{x}) = 1$.

Theorem 1. *The Bayes classifier is given by the equivalent formulas*[1]

$$f^*(\boldsymbol{x}) = \underset{k=1,\ldots,K}{\arg\max}\ \eta_k(\boldsymbol{x}) \tag{1}$$

$$= \underset{k=1,\ldots,K}{\arg\max}\ \pi_k g_k(\boldsymbol{x}). \tag{2}$$

where ties for the maximum may be broken arbitrarily.

Proof. Let $\mathbf{1}_{\{\texttt{pred}\}}$ be 1 if the predicate `pred` is true and 0 otherwise. Then

$$\begin{aligned} R(f) &= \mathbb{E}[\mathbf{1}_{\{f(\boldsymbol{X}) \neq Y\}}] \\ &= \mathbb{E}_{\boldsymbol{X}} \mathbb{E}_{Y|\boldsymbol{X}}[\mathbf{1}_{\{f(\boldsymbol{X}) \neq Y\}}] \\ &= \mathbb{E}_{\boldsymbol{X}} \left[\sum_{k=1}^{K} \eta_k(\boldsymbol{X}) \mathbf{1}_{\{f(\boldsymbol{X}) \neq k\}} \right] \end{aligned}$$

where the second step uses the law of total expectation. In the last expression, for a given $\boldsymbol{X}$, note that all of the indicator functions are 1 except for the one with $k = f(\boldsymbol{X})$. Therefore, to minimize the expression in brackets, f should be such that for each $\boldsymbol{x}$, $f(\boldsymbol{x}) = \arg\max_k \eta_k(\boldsymbol{x})$. This proves (1). Furthermore, this argument works for any f such that $f(\boldsymbol{x})$ outputs a k maximizing $\eta_k(\boldsymbol{x})$. Therefore ties for the maximum can be broken arbitrarily.

To prove (2), note that by Bayes rule,

$$\eta_k(\boldsymbol{x}) = \frac{\pi_k g_k(\boldsymbol{x})}{\sum_{\ell=1}^{K} \pi_\ell g_\ell(\boldsymbol{x})},$$

and that the denominator does not depend on k.

□

3 Plug-in Classifiers

In machine learning, we don't know the joint distribution of $(\boldsymbol{X}, Y)$, and therefore we cannot determine the Bayes classifier. Instead, we observe training data

$$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \overset{iid}{\sim} P_{\boldsymbol{X}Y},$$

where iid stands for *independent and identically distributed*. One approach to classification is to use the training data to estimate either the posterior probability function $\eta_y(\boldsymbol{x})$, or the class-conditional distributions and prior probabilities, and "plug" these estimates in to the corresponding formula for the Bayes classifier.

[1] Technically, the arg max operator returns a set, since multiple values might attain the maximum. In this case, any maximizer can be assigned and still yield an optimal classifier.

We will look at three examples of plug-in classifiers. Each is based on a different assumption about the joint distribution:

- Linear discriminant analysis assumes the class-conditional distributions are multivariate Gaussian with common covariance matrix.
- Naïve Bayes assumes the features X_i are independent when conditioned on the label.
- Logistic regression assumes $\eta_y(\boldsymbol{x})$ is described by a logistic probability model.

Exercises

1. (★) Let f be the constant classifier that always predicts a fixed class k regardless of $\boldsymbol{x}$. Given a simple formula for $R(f)$.

2. (★★) Consider multiclass classification with K classes. As a function of K, what is the largest possible value of the Bayes Risk R^*? Justify your answer.

3. This problem concerns binary classification. For this problem, let the labels be 0 or 1, and define $\eta(\boldsymbol{x}) := \Pr(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$.

   (a) (★★) Let the labels by ±1. Prove that for any classifier f,

   $$R(f) - R^* = \mathbb{E}_{\boldsymbol{X}}[|2\eta(\boldsymbol{X}) - 1|\mathbf{1}_{\{f(\boldsymbol{X}) \neq \text{sign}(2\eta(\boldsymbol{X})-1)\}}]. \tag{3}$$

   The results says that the excess risk depends on how much (on average) $\eta(\boldsymbol{X})$ deviates from $\frac{1}{2}$ at points where h disagrees with the Bayes classifier. *Hint*: Refer to the proof for the Bayes classifier in the lecture notes, and for the binary case rewrite the proof in terms of η.

   (b) (★★) Prove that

   $$R^* = \mathbb{E}_{\boldsymbol{X}}\left[\min\{\eta(\boldsymbol{X}), 1 - \eta(\boldsymbol{X})\}\right].$$

   This result says that the closer $\eta(X)$ is to $\frac{1}{2}$ on average, the larger the Bayes risk.

   (c) (★★★) Let $\alpha \in (0, 1)$, and define the *α-weighted misclassification cost* to be

   $$R_\alpha(f) := \mathbb{E}[(1 - \alpha)\mathbf{1}_{\{f(\boldsymbol{X})=0, Y=1\}} + \alpha\mathbf{1}_{\{f(\boldsymbol{X})=1, Y=0\}}].$$

   This assigns different weights to "false positives" and "false negatives." Note that when $\alpha = \frac{1}{2}$, $2R_\alpha(f) = R(f)$. Determine a formula for the Bayes classifier analogous to (1), and also prove a formula analogous to (3).