Bayes Classifier

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Introduction

We've talked about logistic regression, and discriminant analysis in its geometric version (as was originally introduced by Fisher).
In these slides we present the conceptual framework of Bayes Classifiers.

Main Problem

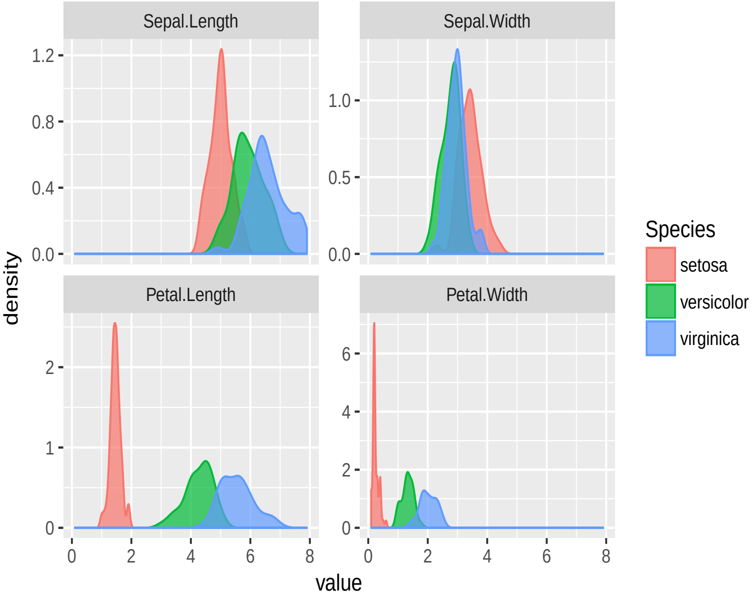
In the learning phase of a classification problem, we typically begin to study the relationship between the predictors and the responses in the form of $(Y|X)$.

That is, for a given group k, we examine how the values of predictors $X_1, X_2, \ldots, X_p$ are distributed.

Dataset iris in R

```
head(iris)

  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```


Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
density
value
1.2
0.8
0.4
0.0
1.0
0.5
0.0
2
1
0
6
4
2
0
0
2
4
6
8
0
2
4
6
8
Species
setosa
versicolor
virginica

Main Problem

In the classification (or decision) phase, what we are interested in is not $P(X|Y)$, but $P(Y|X)$

That is, given the predictor values of an unclassified object, we want to know to which class we should assign the object to.

Key Question

Thus, the crux of the matter consists of using the observed information in $P(X|Y)$ to find $P(Y|X)$.

How do you do that?

Key Question

Thus, the crux of the matter consists of using the observed information in $P(X|Y)$ to find $P(Y|X)$.

How do you do that?

Bayes Theorem to the rescue!

Bayes Theorem

Recall that Bayes theorem (in its general form) says:

$$Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k)Pr(Y = k)}{Pr(X = x)}$$

where $Pr(x)$ is calculated with the total proability formula:

$$Pr(X = x) = \sum_{k} Pr(X = x|Y = k)Pr(Y = k)$$

Bayes Theorem

We can use Bayes Theorem for classification purposes, changing some of the notation:

- $Pr(Y = k) = \pi_k$, the **prior** probability for class k.
- $Pr(X = x|Y = k) = f_k(x)$, the **density** for X in class k.

$$Pr(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{k=1}^{K} f_k(x)\pi_k}$$

Bayes Rule

By using Bayes Theorem we are essentially modeling the posterior probability $P(Y = k|X = x)$ in terms of densities $f_k(x)$ and prior probabilities π_k.

Under this mindset, it seems reasonable to classify an object x_0 to the class k that renders $P(Y = k|X = x_0)$ maximum. That is, classify x_0 to the most likely class, given its predictors.

Formal Framework

Let's formalize things

- We will place ourselves in territory of random variables and probability spaces.
- I will consider one predictor X and one response Y (although things can be generalized to multiple predictors).
- This framework involves some concepts from Statistical Decision Theory.

A bit of Decision Theory

- Let $X \in \mathbb{R}$ denote a real valued random input variable.
  Actually X does not have to be necessarily real; it can be qualitative
- Let G be a set of discrete values (i.e. the classes) with $K = card(\mathsf{G})$.
- A response variable Y takes discrete values in G.
- Let $Pr(Y, X)$ be the joint distribution.
- We seek an estimate $\hat{Y} = f(X)$ for predicting Y given the values of the input X.
- The estimate $\hat{Y}$ will assume values in G.

Loss Function

This theory requires a **loss function** for penalizing errors in prediction:

$$Loss(Y, f(X))$$

- The loss function for classification tasks is represented by a $K \times K$ matrix $\mathbf{L}$.
- This matrix will be zero on the diagonal and nonnegative elsewhere.
- The element $L_{k,l}$ in k-th row and l-th column is the price paid for classifying an observation belonging to class G_k as G_l.
- Most often we use the *zero-one* cost, where all misclassifications are charged one unit.

Expected Prediction Error

The criterion for choosing $\hat{Y}(X)$ is the so-called **Expected Prediction Error** (EPE):

$$\mathsf{EPE}(f) = E\left[Loss(Y, \hat{Y}(X))\right]$$

where the expectation is taken with respect to the joint distribution $Pr(Y, X)$.

Expected Prediction Error

Taking the expectation with respect to the joint distribution $Pr(Y, X)$

$$\mathsf{EPE}(f) = E_X \left\{ \sum_{k=1}^{K} L(G_k, \hat{Y}(X)) Pr(G_k|X) \right\}$$

And we look for $f()$ that minimizes EPE.

Expected Prediction Error

It suffices to minimize EPE pointwise:

$$\hat{Y}(X) = \underset{g \in \mathsf{G}}{argmin} \left\{ \sum_{k=1}^{K} L(G_k, g) Pr(G_k | X = x) \right\}$$

With the 0-1 loss function this simplifies to:

$$\hat{Y}(X) = G_k \quad \text{if} \quad Pr(G_k | X = x) = \underset{g \in \mathsf{G}}{max} Pr(g | X = x)$$

Bayes Classifier

$$\hat{Y}(X) = G_k \quad \text{if} \quad Pr(G_k|X = x) = \max_{g \in \mathsf{G}} \{Pr(g|X = x)\}$$

- Is known as the **Bayes Classifier** (or Bayes Rule)
- It says that we classify to the most probable class, using the conditional distribution $Pr(Y|X)$
- It produces the lowest possible error rate, called the *Bayes error rate*.

Summary: Bayes Classifier

Decision theory tells us that we need to know the class posteriors $Pr(Y = k|X)$ for optimal classification.

$$Pr(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{k=1}^{K} f_k(x)\pi_k}$$

It does make sense to use the formula from the Bayes theorem for classification purposes.

Wrapping things up

Keep in mind

The Bayes formula is "the way to go"

$$Pr(Y = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{k=1}^{K} f_k(x)\pi_k}$$

in the sense that we should assign each observation to the most likely class, given its predictor values.

Keep in mind

However, the Bayes formula

$$Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

does NOT tell us:

- how to calculate priors π_k
- what form should we use for densities $f_k(x)$

There is plenty of room to play with π_k and $f_k(x)$

Open Questions

How do we estimate priors π_k?

What density $f_k(x)$ do we use?

- Normal distribution(s)?
- Mixture of Normal distributions?
- Non-parametric estimates (e.g. kernel densities)?
- Assume predictors are independet (Naive Bayes)?

Keep in mind that a Bayes Classfier works as long as the terms in $Pr(Y = k|X = x)$ are all correctly specified.

Open Questions

Interestingly, we can also try to directly specify the posterior $Pr(Y = k|X)$ with a *semi-parametric* approach, for instance:

$$Pr(Y = k|X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

If we choose this approach, is this still optimal?
(i.e. can this be a Bayes Classifier?)

Bibliography

- **The Elements of Statistical Learning** by Hastie et al (2009). *Chapter 2, section 2.4: Statistical Decision Theory*. Springer.
- **An Introduction to Statistical Learning** by James et al (2013). *Chapter 2, section 2.2.3: The Classification Setting*. Springer.