Classification Basics

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

Introduction

The goal in classification is to take an input vector $\mathbf{x}$ and to assign it to one of K discrete classes or groups G_k where $k = 1, \ldots, K$.

In the most common case, the classes are taken to be disjoint, so that each input is assigned to one and only one class.

Credit Score Example

2 Classes

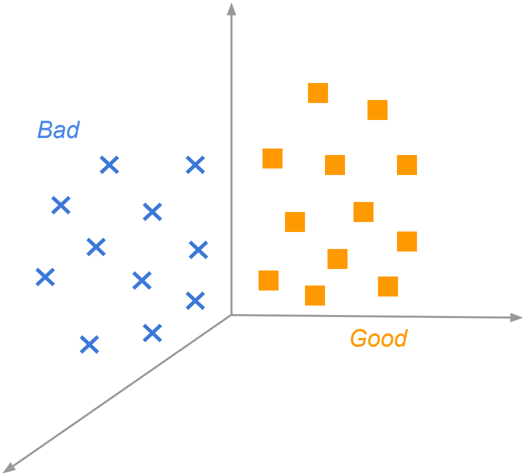
- Let's consider a credit application from which p predictors are derived $X = [X_1, \ldots, X_p]$.
  - Age
  - Job type (and job seniority)
  - Residential Status
  - Marital Status
  - Loan purpose
  - *etc*

- Customers are divided in two classes: "good" and "bad"
  - Good customers are those that payed their loan back
  - Bad customers are those that defaulted on their loan

Classification Idea

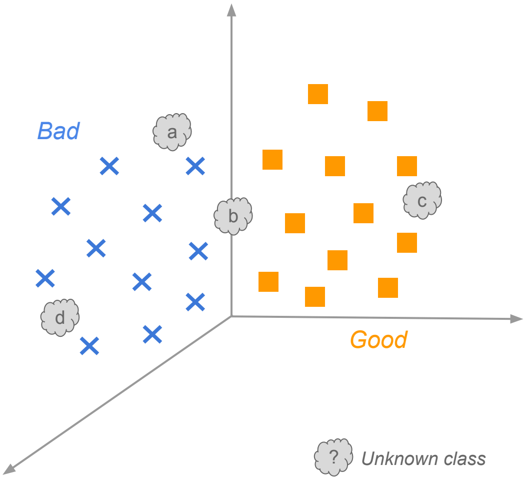
- Given a customer's attributes $X = x$, to what class Y we should assign this customer?
- Ideally (although not mandatory), we would like to know: $P(Y|X = x)$

Data set of good and bad customers



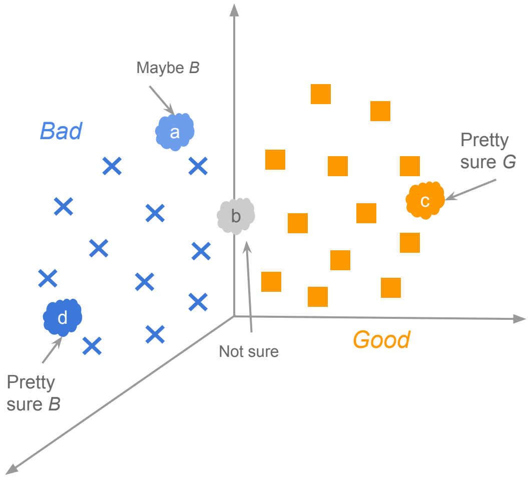
Cloud of n points in p-dimensional space

How to classify new customers?



To which class we assign new individuals?

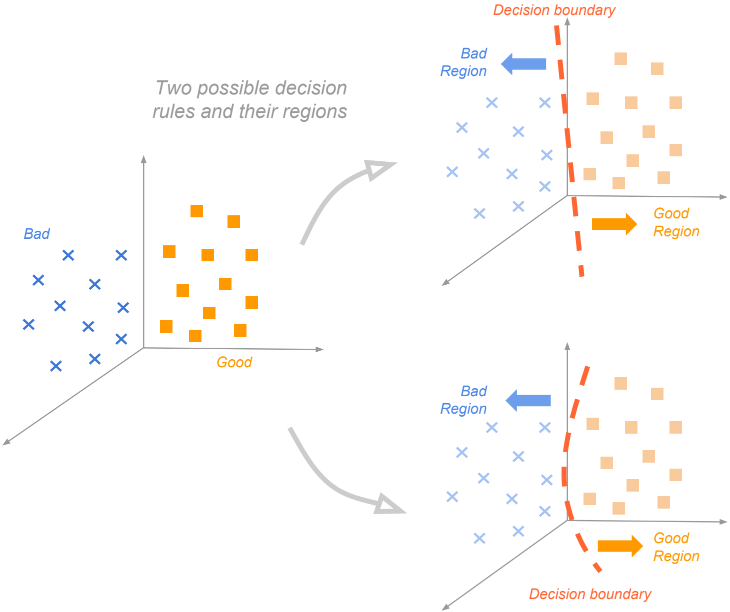
How to classify new customers?



Some possible classifications

Classification and Decisions

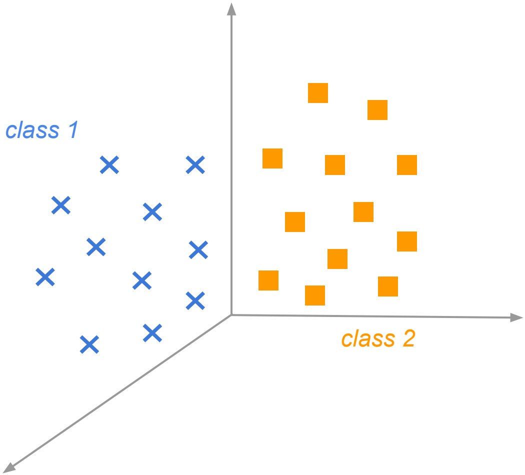
- It would be nice to have a mechanism or **rule** to classifiy observations (i.e. to make a decision).
- Such a rule would divide the input space into regions R_k called **decision regions** (one for each class).
- The boundaries between decision regions would establish **decision boundaries** (or decision surfaces).
- We are going to study different approaches to determine such decision rules.



Two possible decision rules and their regions
Bad
Good
Decision boundary
Bad Region
Good Region
Bad Region
Good Region
Decision boundary

Motivation

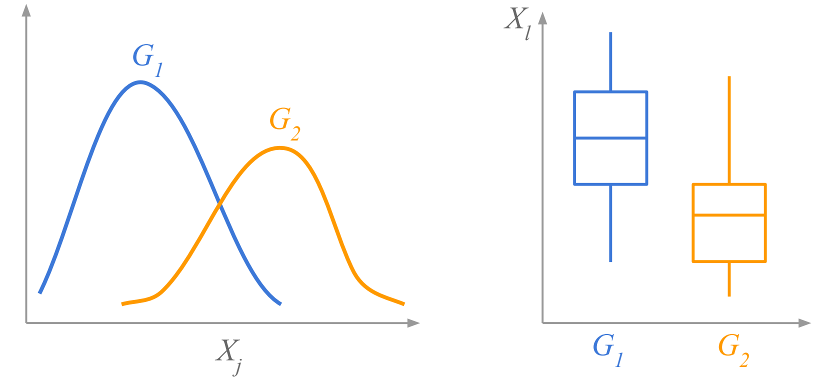
Let's consider a two class problem



Cloud of n points in p-dimensional space

Let's consider a two class problem

- Since this is a supervised learning problem, we start with some available evidence (i.e. training data set)
- A first step may involve studying how X values vary according to a given class k
- In other words, we may start exploring the conditional distribution of $X = x|Y = k$
- e.g. How does $X_j|Y = 1$ compare with $X_j|Y = 2$?



Exploring conditional distributions $X|Y = k$

Conditional Probabilities

- The main challenge is about making (accurate) predictions of a “new” individual.
- How to guess the class of an observation $\mathbf{x_0}$?
- Typically we have information about $X|Y = k$
- Often, we may even be able to guesstimate $P(X|Y = k)$
- But what we need is $P(Y = k|X = x)$

Conditional Probabilities

Pretend we know the *population* distribution $P(X = x|Y = k)$

$$P(X = x|\text{Good}) = \frac{P(\text{applicant is Good and has attributes } x)}{P(\text{applicant is Good})}$$

similarly

$$P(X = x|\text{Bad}) = \frac{P(\text{applicant is Bad and has attributes } x)}{P(\text{applicant is Bad})}$$

Conditional Probabilities

But what we really want is $P(Y = k|X = x)$

$$P(\text{Good}|X = x) = \frac{P(\text{applicant has attributes } x \text{ and is Good})}{P(\text{applicant has attributes } x)}$$

similarly

$$P(\text{Bad}|X = x) = \frac{P(\text{applicant has attributes } x \text{ and is Bad})}{P(\text{applicant has attributes } x)}$$

Conditional Probabilities

Is there a connection between:

$$P(X = x|Y = k) \qquad \text{and} \qquad P(Y = k|X = x)\ ?$$

Conditional Probabilities

Is there a connection between:

$$P(X = x|Y = k) \qquad \text{and} \qquad P(Y = k|X = x) \;?$$

YES!!!

Bayes' Rule

Bayes' Rule Reminder

Bayes' Rule

Let's look at both types of conditional probabilities:

$$P(X = x|Y = k) = \frac{P(Y = k \text{ and } X = x)}{P(Y = k)}$$

and

$$P(Y = k|X = x) = \frac{P(X = x \text{ and } Y = k)}{P(Y = k)}$$

solving for $P(X = x \text{ and } Y = k)$ we have that:

Bayes' Rule

Solving for $P(X = x$ and $Y = k)$ we have that:

$$P(X = x|Y = k)P(Y = k) \; = \; P(Y = k|X = x)P(X = x)$$

Thus:

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

Bayes Theorem

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

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

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

$$P(X = x) = \sum_k P(X = x|Y = k)P(Y = k)$$

Bayes Theorem

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

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

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

Bayes Rule

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

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

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.

Classification Error

Bayes' Rule

Assuming that we know $P(Y = k|X = x)$, we can use it to make two guesses:

- guess Good with $P(\text{Good}|X = x)$
- guess Bad with $P(\text{Bad}|X = x)$

Which class should we choose? Assign applicant to class k for which $P(Y = k|X = x)$ is the largest.

This seems like a reasonable idea... but is it optimal?

Classification decision

In a two-class problem, Whenever we observe a particular x, we can make four decisions:

- guess "Good" when the applicant is Good (correct decision)
- guess "Good" when the applicant is Bad (incorrect decision)
- guess "Bad" when the applicant is Bad (correct decision)
- guess "Bad" when the applicant is Good (incorrect decision)

Confusion Matrix

		Reality	
		Good	Bad
Guess	Good	a	b
	Bad	c	d

$$n = a + b + c + d$$

Confusion Matrix

		Reality	
		Good	Bad
Guess	Good	a	b
	Bad	c	d

Correct classifications

Incorrect classifications

$$n = a + b + c + d$$

Classification Rates

From the previous 2×2 confusion matrix we can obtain two types of classification rates:

$$\text{Correct classification rate } = \frac{a+d}{n}$$

$$\text{Misclassification rate } = \frac{b+c}{n}$$

It seems reasonable to obtain a correct classfication rate as large as possible, or conversely, minimize the error classification rate.

Probability of error

Whenever we observe a particular x, the probability of error is:

$$P(error|x) = \begin{cases} P(\text{Good}|X = x) & \text{if we decide Bad} \\ P(\text{Bad}|X = x) & \text{if we decide Good} \end{cases}$$

Probability of error

For a given x we can minimize the probability of error by deciding "Good" if $P(\text{Good}|X = x) > P(\text{Bad}|X = x)$, and "Bad" otherwise.

However, we may never observe exactly the same value of x twice. Will this rule minimize the probability of error?

Yes, because the average probability of error is given by:

$$\int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error|x)P(x)dx$$

If for every x we assure that $P(error|x)$ is as small as possible, then the previous integral must be as small as possible.

Probability of error

The *Bayes decision rule* for minimizing the probability of error:

$$\text{Decide “Good” if } P(\mathsf{Good}|X = x) > P(\mathsf{Bad}|X = x);$$
$$\text{othwersize decide “Bad”}$$

becomes

$$P(error|x) = min\{P(\mathsf{Good}|X = x),\ P(\mathsf{Bad}|X = x)\}$$

Probability of error

Later on we will formalize the ideas in these slides, and generalize them in a couple of ways:

- by allowing more than two classes
- by looking at ways to estimate all the required probabilities and densities
- by introducing a *loss function* more general than the probability of error

Wrapping things up

Keep in mind

The Bayes formula is "the way to go"

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

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

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

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 $P(Y = k|X)$ with a *semi-parametric* approach, for instance:

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

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.
- **Pattern Recognition and Machine Learning** by Christopher Bishop (2006) *Chapter 1: Introduction*. Springer.
- **Statistical Regression and Classification** by Norman Matloff (2017) *Chapter 1: Setting the Stage*. CRC Press.
- **Pattern Classification** by Duda, Hart, and Stork (2006). *Chapter 2: Bayesian Decision Theory*. Wiley.

French Literature

- **Modeles Statistiques pour Donnees Qualitatives** by Droesbeke et al (2005). *Chapter 6: Modele a reponse dichotomique* by P.L. Gonzalez. Editions Technip, Paris.
- **Statistique Explicative Appliquee** by Nakache and Confais (2003). *Chapter 4: Modele logistique binaire*. Editions Technip, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 18: Analyse discriminante et regression logistique*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 11: La regression logistique binaire*. Dunod, Paris.