Introduction to Statistical Learning

Week 2

Daniele Bianchi[1]

whitesphd.com

[1]School of Economics and Finance
Queen Mary, University of London

Summary

This week we introduce basic concepts of statistical learning, such as predictive accuracy and the bias-variance trade-off. In addition, we will cover basics of classification methods with examples from financial markets.

Contents

Statistical Learning

What is Statistical Learning?

Take the functional relationship

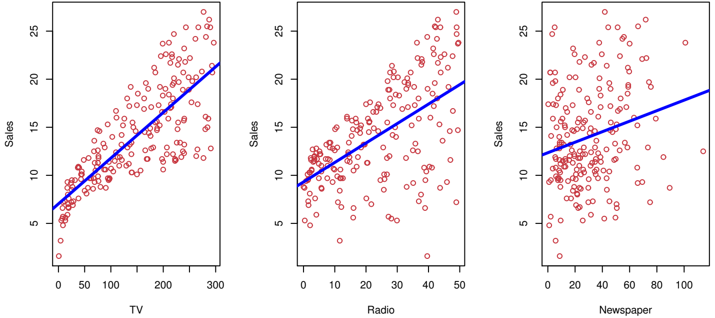
$$Y = f(X) + \epsilon,$$

With

- Y an observed quantitative response
- $X = (X_1, X_2, \ldots, X_p)$ a set of p predictors
- f an *unknown* function of X
- ϵ is the *error term*, independent of X, and with mean zero.

f represents the *systematic information* that X provides about Y.

Example: Advertising



Sales vs TV, Radio and Newspaper, with a blue linear-regression line fit separately to each. Source: "An introduction to Statistical Learning", James et al. (2013).

Can we predict Sales using these linear models? Perhaps we can do better:

$$Sales \approx f\left(TV, Radio, Newspaper\right),$$

What is $f(X)$ good for?

Prediction: With a good f we can make predictions of Y at new points $X = x$.

Key question: "is there an ideal $f(X)$? That is, what is a good value for $f(X)$ at any selected value of X, say $X = 4$?

Suppose you have a $\hat{f}$, which predicts, Y, i.e., $\hat{Y}$:

$$\hat{Y} = \hat{f}(X),$$

Assume that both $\hat{f}$ and X are fixed for now. Then,

$$\begin{aligned} E\left(Y - \hat{Y}\right)^2 &= E\left[f(X) + \epsilon - \hat{f}(X)\right]^2, \\ &= \underbrace{\left[f(X) - \hat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}}, \end{aligned}$$

Why estimate f?

So the goal here is to estimate f with the aim of reducing the "reducible" error, as much as possible.

Notice there is not much we can do about the "irreducible" error $Var\left(\epsilon\right)$

In fact, even if we knew $f(X)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values, i.e., there is "uncertainty" around X.

In this lecture, we cover two alternative methods to estimate f.

How do we estimate f?

Two classes of methods:

- **Parametric methods:** Model-based approach (Example: Linear model, Ordinary Least Squares method)
- **Non-parametric methods:** No explicit assumptions about the functional form f. Learn from the data.

Each classes of methods has its own pros and cons. Out goal today is to understand some of these methods under each of these approaches.

Spoiler: Parametric methods are less flexible but easier to interpret, while non-parametric methods are more flexible but harder to interpret.

Example: Income and education

Suppose we have to estimate the relationship between Income, Education and Seniority,

$$income = f\left(education, seniority\right) + \epsilon,$$

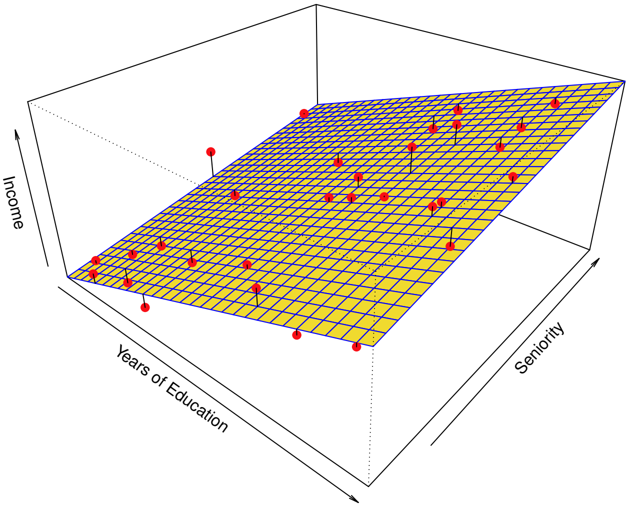
We estimate two alternative models:

- A linear regression model

$$\hat{f}\left(education, seniority\right) = \beta_0 + \beta_1 education + \beta_2 seniority,$$

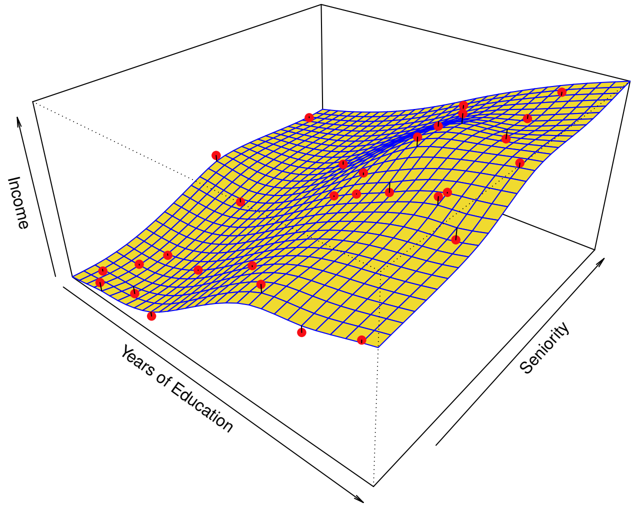
- A non-parametric spline regression (more later in the course)

Example: Income and education



This figure shows the mapping between income, education and seniority based on a parametric linear model. Source: "An introduction to Statistical Learning", James et al. (2013).

Example: Income and education



This figure shows the mapping between income, education and seniority based on a non-parametric splines linear model. Source: "An introduction to Statistical Learning", James et al. (2013).

Model accuracy and trade-offs

Some trade-offs

Prediction accuracy versus interpretability.

- Linear models are easy to interpret; spline regressions are not.

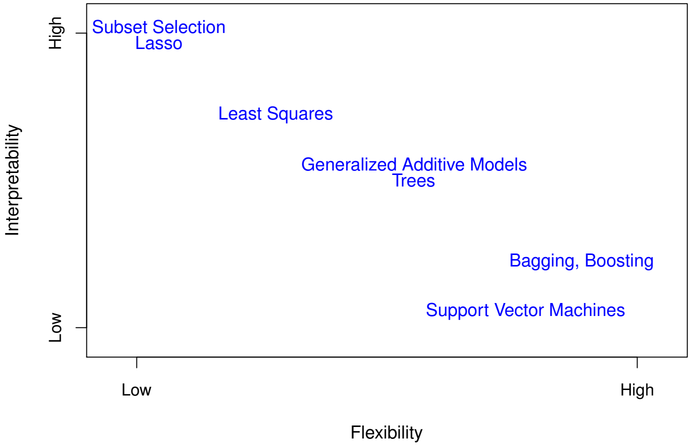
Good fit versus over-fit or under-fit.

- How do we know when the fit is just right?

Parsimony versus black-box.

- We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.

Prediction accuracy versus interpretability



This figure shows the trade-off between interpretability and prediction accuracy for different statistical learning techniques. Source: "An introduction to Statistical Learning", James et al. (2013).

Assessing model accuracy

When it comes to choosing the right model few key things should be considered:

- No method is the best for *all* settings. Methods are often "application-specific"..
- Different settings demand different approaches
- If that is the case, how does one select a model?
- Goal: Quantify the extent to which our prediction from a model is close to the true response.

Within the context of regression models, such "closeness to the truth" is measure by the so-called Mean Squared Error (MSE).

Model accuracy: Regression problem

The Mean Squared Error (MSE) is defined as,

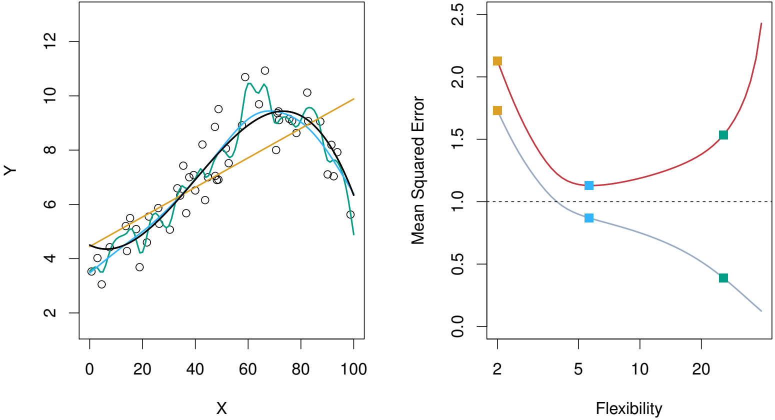
$$\mathsf{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

- Most commonly used measure in a regression setting.
- Result: MSE will be very small if the predicted responses are very close to the true responses (large, otherwise).
- But we can overfit the sample! You may be able to find a model that minimises the MSE in the *training data*.
- However, need not mean that new incoming information about X will do just as good a job of predicting Y.

Model accuracy: Regression problem

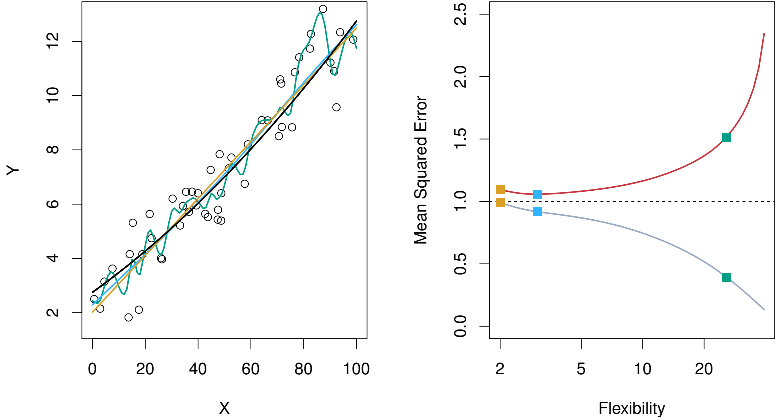
- Suppose we want to predict stock returns using past returns.
- We can train the model with data from the past year, and check MSE.
- But we don't care about whether the model does a good job of predicting the past.
- What is most useful is how much of the return *tomorrow* can the model predict?
- We want to train the model with data, and compute MSE with data previously unseen by the model, i.e., test data
- Key approach to evaluate how well your $\hat{f}$ is doing.

Test MSE, Degrees of Freedom



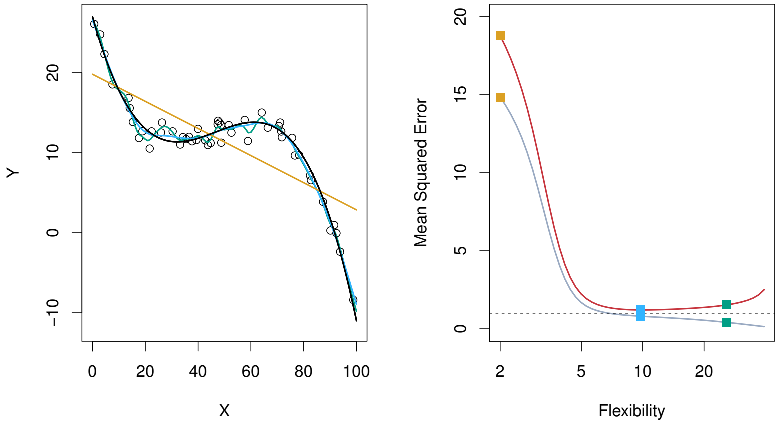
Notes: Circles are the data. Black curve is truth f. Training MSE (Grey curve), and Test MSE (red curve). Left graph: Models. Source: "An introduction to Statistical Learning", James et al. (2013).

Test MSE, Degrees of Freedom



Notes: Circles are the data. Black curve is truth f. Training MSE (Grey curve), and Test MSE (red curve). Left graph: Models. Source: "An introduction to Statistical Learning", James et al. (2013).

Test MSE, Degrees of Freedom



Notes: Circles are the data. Black curve is truth f. Training MSE (Grey curve), and Test MSE (red curve). Left graph: Models. Source: "An introduction to Statistical Learning", James et al. (2013).

Bias-Variance trade-off

- Turns out that U shape observed in the graph is a result of competing properties of any learning method.
- A given value of test MSE can be decomposed into:
    1. Variance of $\hat{f}(x_0)$.
    2. Squared bias of $\hat{f}(x_0)$.
    3. Variance of the error term, $\text{Var}(\epsilon)$.

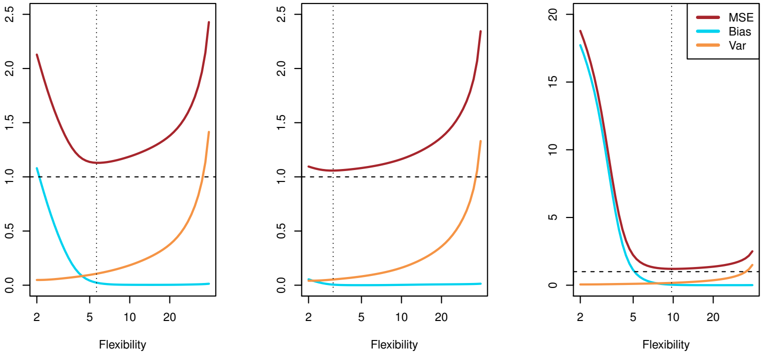
$$E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon),$$

Bias-Variance trade-off

$$E(y_0 - \hat{f}(x_0))^2 = \mathsf{Var}(\hat{f}(x_0)) + [\mathsf{Bias}(\hat{f}(x_0))]^2 + \mathsf{Var}(\epsilon),$$

- Variance: Amount by which $\hat{f}$ would change if we used a different training data set.
- Bias: Estimating a very complex real world with a simple model. If real model f is non-linear (unobserved), any linear model $\hat{f}$ will have high bias.
- Typically as the flexibility of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a **bias-variance trade-off**.

Bias-Variance trade-off



Source: "An introduction to Statistical Learning", James et al. (2013).

Classification

Model accuracy: Classification

Suppose your Y is no longer a continuous variable, but is **qualitative**, e.g., a firm is defaulted or not.

Our goal is no longer build the best conditional expectation, such as in regression problems, but is:

- Build a classifier $C(X)$ that assigns a class label from $\mathcal{C}$ to a future unlabeled observation X.
- Assess the uncertainty in each classification.
- Understand the role of different predictors among $X = (X_1, X_2, \ldots, X_p)$.

Model accuracy: Classification

Typically, we measure the performance of the classifier $\hat{C}(X)$ using the classification (or training) error rate.

We compute the *training error* rate as:

$$\frac{1}{n}\sum_{i=1}^{n} I\left(y_i \neq \hat{y}_i\right),$$

i.e., have we correctly "classified" using our model/method?

Just as with the MSE, we can also compute the *test error rate* associated with a set of test observations.

What fraction of the test data is wrongly classified?

The Bayes Classifier

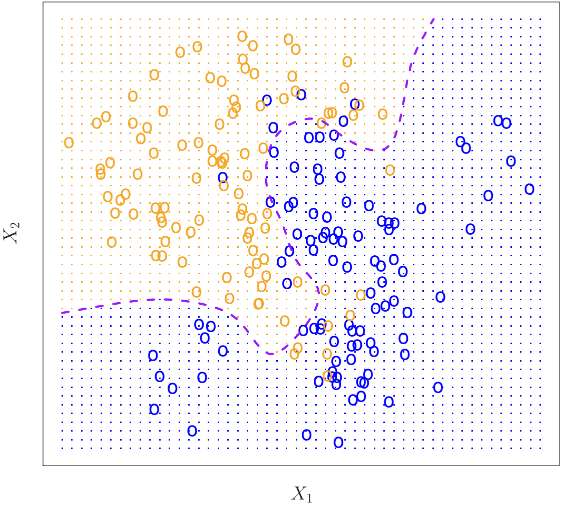
In classification problems we want to minimize the test error rate.

A simple approach that assigns each observation to the most likely class *given its predictor values* achieved this, i.e.,

$$\mathsf{Pr}\left(Y = j | X = x_0\right),$$

This very simple approach is called the "Bayes classifier".

The Bayes Classifier: Example



Orange shade: $\Pr(Y = \text{orange}|X) > 50\%$. Blue shade: Otherwise. Dashed line: Exactly 50% (Decision boundary). Source: "An introduction to Statistical Learning", James et al. (2013).

The Bayes Classifier: Error Rate

It can be shown (beyond this course) that the Bayes classifier produces, theoretically, the lowest possible test error rate.

This lowest error rate is known as the *Bayes error rate*.

The Bayes classifier will always choose the class for which the conditional probability $\Pr(Y = j|X = x_0)$ is the largest.

Error rate at $X = x_0$ is $1 - E(\max \Pr(Y = j|X = x_0))$.

K-nearest Neighbours (KNN) classifier

The Bayes classifier is ideal, but only theoretically, not for real data. In real applications you often have no idea what the conditional distribution of $Y|X$ is.

That is, in reality the Bayes classifier is a "gold standard", i.e., unattainable in practice.

A feasible approach is:

1. Estimate the conditional distribution of $Y|X$
2. Classify an observation to the class with *highest estimated probability*.

One method to do this is the K-nearest neighbours (KNN) classifier.

K-nearest Neighbours (KNN) classifier

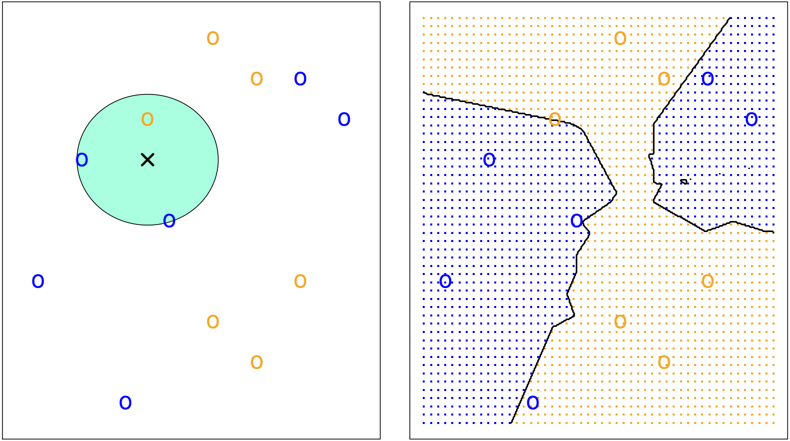
Given a positive integer K, and a test observation x_0,

- Identify K points that are closest to x_0, i.e., $\mathcal{N}_0$.
- Estimate conditional probability for class j as a fraction of points in $\mathcal{N}_0$ whose response value equal j.

$$\Pr\left(Y = j | X = x_0\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I\left(y_i = j\right).$$

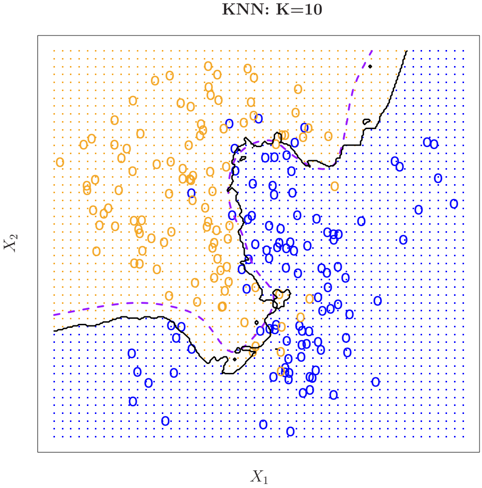
- KNN applies Bayes rule and classifies the observation x_0 to the class with the largest probability.

K-nearest Neighbours (KNN) classifier



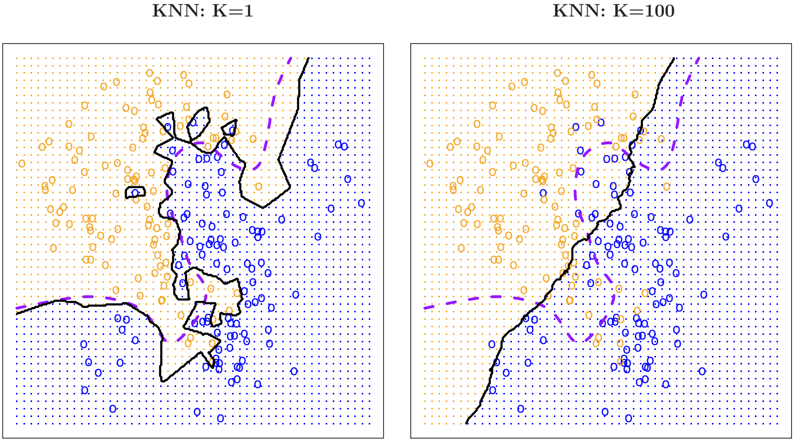
Source: "An introduction to Statistical Learning", James et al. (2013).

K-nearest Neighbours (KNN) classifier



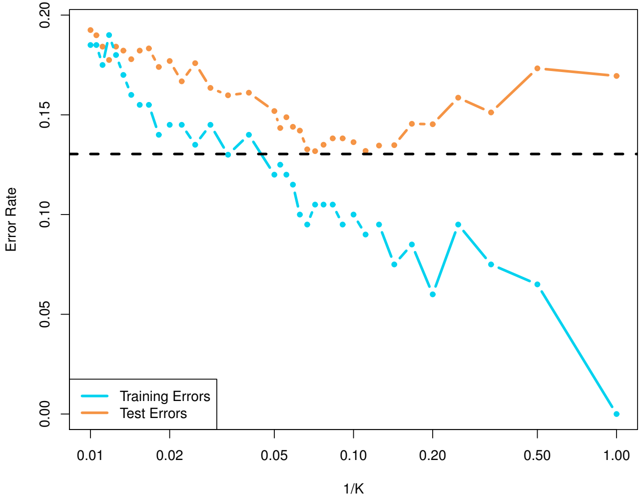
Source: "An introduction to Statistical Learning", James et al. (2013).

K-nearest Neighbours (KNN) classifier



Source: "An introduction to Statistical Learning", James et al. (2013).

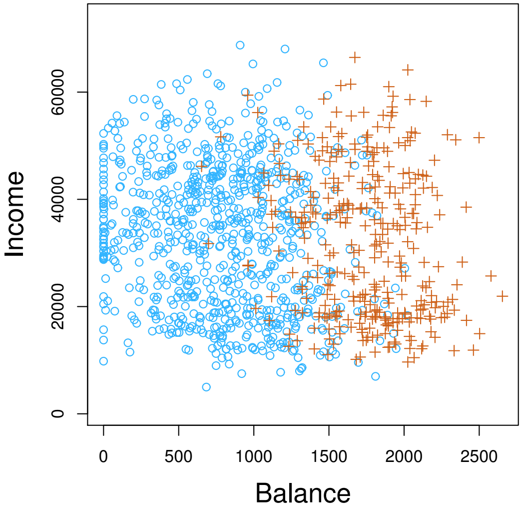
K-nearest Neighbours (KNN) classifier



Source: "An introduction to Statistical Learning", James et al. (2013).

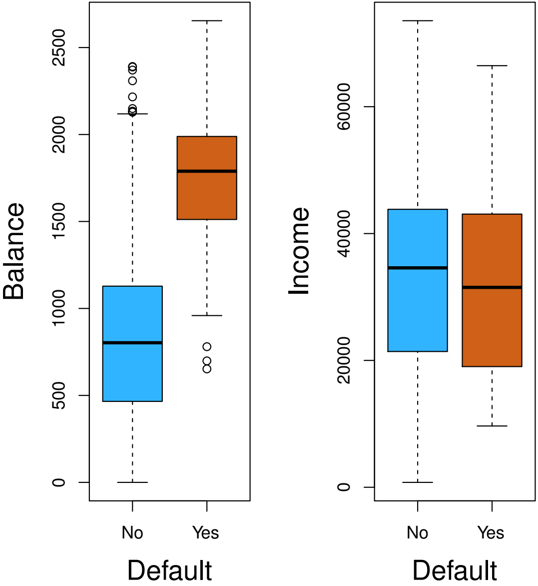
Linear Discriminant Analysis

Why classification?

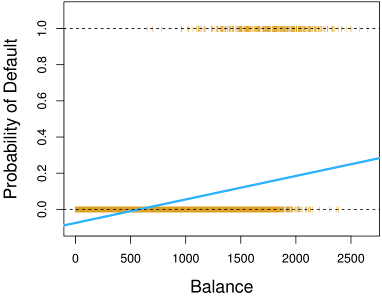


Source: "An introduction to Statistical Learning", James et al. (2013).

Why classification?



Why classification?



Source: "An introduction to Statistical Learning", James et al. (2013).

Logistic Regression

- Linear regression: $p(X) = \beta_0 + \beta_1 X$
- For balances close to zero, we predict a *negative* probability of default. Unhelpful, not sensible.
- We have to model $p(X)$ using a function that gives outputs between 0 and 1 for all values of X.
- Many different functions available. But we use the logistic function:

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}},$$

- The quantity of interest (after manipulating the equation above) is:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X},$$

- This is known as the *odds ratio*. A $p(X) = 0.2$, implies an odds of $1/4$.

Linear Discriminant Analysis

- Logistic regression models Pr$(Y = k|X = x)$.
- Alternative approach:
    - Model the distribution of predictors X for each response class k.
    - Then use Bayes theorem to estimate Pr$(Y = k|X = x)$.
- Logistic regressions are unstable when the k classes of Y are well-separated.

Linear Discriminant Analysis

- Bayes Theorem states:

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

- $f_k(x) = Pr(X = x|Y = k) \to$ probability density function.
- $\pi_k \to$ prior probability that a randomly chosen observation belongs to category k.
- Estimate π_k and $f_k(X)$ separately and then use Bayes theorem.

Linear Discriminant Analysis

- Suppose we have one predictor.
- Assumption: the density function $f_k(x)$ is normally distributed.
- The Linear Discriminant Analysis (LDA) classifier, uses this assumption to estimate π_k and $f_k(X)$ and then classify the observations into different categories k.

Linear Discriminant Analysis: Error rates

Confusion matrix

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
Default	Yes	23	81	104
Status	Total	9,667	333	10,000

- Training error rate = 2.75%. (275 misclassified).
- Only 3.3% of the individuals in the training sample defaulted.
- The algorithm that does not use additional variables will always predict with an error rate of 3.3%.
- LDA doesn't do that much better, i.e., the additional explanatory variables only improved error rate to 2.75%.

Can we do better?

- A firm (credit card company / bank) will be interested to reduce false negatives, i.e.,
- Avoid incorrectly classifying an individual who will default!
- Reverting to the Bayes classifier:

$$Pr(\text{default} = \text{Yes}|X = x) > 0.5$$

- LDA uses the same threshold of 50%. We can consider lowering the threshold!
- Set the problem to be:

$$Pr(\text{default} = \text{Yes}|X = x) > 0.2$$

Can we do better?

$Pr(\text{default} = \text{Yes}|X = x) > 0.5$

		True default status		
		No	Yes	Total
Predicted	No	9,644	252	9,896
Default	Yes	23	81	104
Status	Total	9,667	333	10,000

$Pr(\text{default} = \text{Yes}|X = x) > 0.2$

		True default status		
		No	Yes	Total
Predicted	No	9,432	138	9,570
Default	Yes	235	195	430
Status	Total	9,667	333	10,000

Wrap-up

Summary

- We looked at new methods beyond linear regression to estimate $\hat{f}$.
- New tools to evaluate how to assess whether the model describes the data well.
- Test vs. Training error rates and MSE.
- The following set of lectures will use all of these tools in Finance.
- The classes will help you learn how to estimate these models with data.