QBUS6810: Statistical Learning and Data Mining

Lecture 1: Introduction to Statistical Learning

Semester 2, 2018

Discipline of Business Analytics, The University of Sydney Business School

Lecture 1: Introduction to Statistical Learning

1. Introduction
2. Business Examples
3. Notation
4. Statistical decision theory
5. Evaluating performance
6. Overview of some key concepts

Introduction

Introduction

Machine learning is a set of methods for automatically detecting patterns in data and using them for predicting future data and guiding decision making. In other words, learning from data.

We can think of **statistical learning** as a framework for machine learning that draws on statistics.

There are various reasons for following this approach, such as:

- Studying the statistical properties of learning methods.
- Drawing on insights from statistics to develop learning methods.
- Generating probabilistic predictions and quantifying uncertainty.

Introduction

Two trends bring statistical learning to the forefront of successful business decision making:

- We are in the era of **big data**. The Internet and increasing presence of data capturing devices (such as mobile phones, cameras, sensors, card readers, etc), combined with large reductions in the cost of storage, brought an unprecedented availability of data, and continued dramatic growth in the size of data sets.

- Advances in computing power increase the scope for exploring complex patterns in data.

Types of Learning

There are two main types of learning:

- In **predictive** or **supervised learning**, the objective is to learn a function to predict an output variable Y based on observed input variables $X_1, \dots, X_p$.

- In **descriptive** or **unsupervised learning**, we only have inputs, $X_1, \dots, X_p$, and the goal is to find "interesting" patterns in this data.

Supervised learning

In supervised learning, the output, or **response variable**, can be of any type. However, most methods address two main classes of supervised learning problems:

- In **regression**, the response is quantitative (such as a person's credit card balance).
- In **classification**, the response is a categorical variable. The case where the response takes only 2 different values corresponds to binary classification.

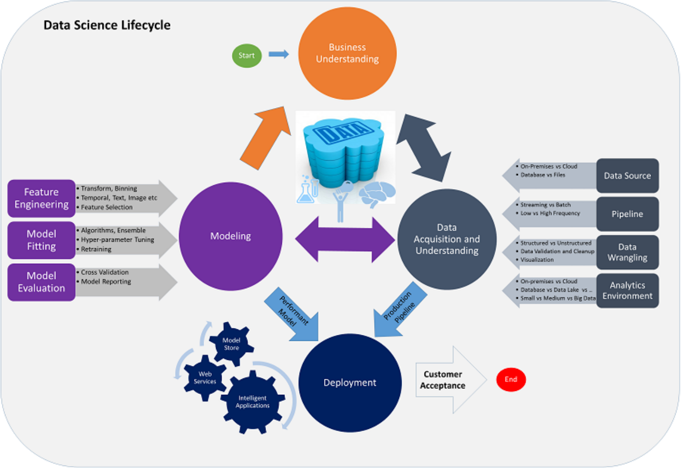
Data mining and data science

Data mining is the process of extracting interesting and previously unknown patterns and relations from large databases, drawing on the fields of machine learning, statistics, and database technology.

Data science is a multidisciplinary field that combines knowledge and skills from statistics, machine learning, software engineering, data visualisation, and domain expertise (in our case, business expertise) to uncover value from large and diverse data sets.

Data scientists often work directly with stakeholders (say, product managers) to translate data analysis results into action.

The data science process: a real-world perspective



https://docs.microsoft.com/en-us/azure/machine-learning/data-science-process-overview

Learning outcomes

After successfully completing this unit you are expected to:

1. Understand the conceptual and theoretical foundations of statistical learning.
2. Develop an in-depth knowledge of regression and classification learning methods for business applications.
3. Be able to conduct a complete data analysis project based on these foundations and methods.
4. Know how to use Python for your practical workflow under realistic data complexity.
5. Effectively communicate your results to guide decision making.

Business Examples

Zillow Kaggle competition

- Kaggle is a crowdsourcing platform that allows organisations to post data prediction problems to be solved through public competitions.
- Zillow's Home Value Prediction is a recent competition (with a 1.2 million dollar cash prize) that invited participants to make predictions about the future sale prices of homes (a regression problem).
- In this competition, the goal was to improve on Zillow's home valuation estimates ("ZEstimates"), which are based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property.

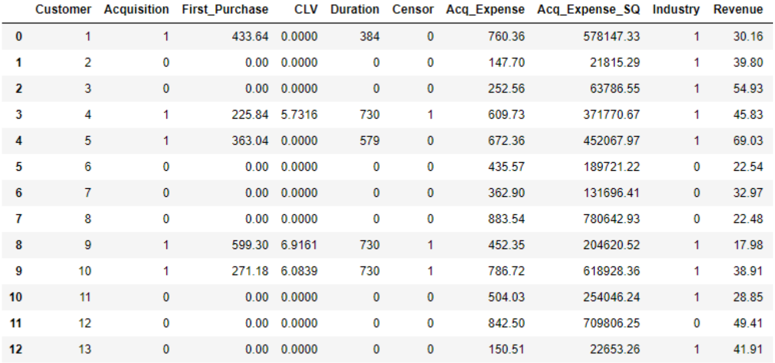
Customer relationship management

- Customer relationship management (CRM) is a set of practices on collecting and studying customer information with the objective of maximising customer lifetime value (CLV), the net value of a customer to a firm over his/her entire lifetime.

- CRM can be part of a customer-centric (as opposed to brand-centric) business strategy, that focuses on the acquisition and retention of profitable customers.

- CRM has four main areas: customer acquisition, retention, churn, and win-back. Statistical models and machine learning algorithms play a central role in each of these areas.

Customer relationship management



	Customer	Acquisition	First_Purchase	CLV	Duration	Censor	Acq_Expense	Acq_Expense_SQ	Industry	Revenue
0	1	1	433.64	0.0000	384	0	760.36	578147.33	1	30.16
1	2	0	0.00	0.0000	0	0	147.70	21815.29	1	39.80
2	3	0	0.00	0.0000	0	0	252.56	63786.55	1	54.93
3	4	1	225.84	5.7316	730	1	609.73	371770.67	1	45.83
4	5	1	363.04	0.0000	579	0	672.36	452067.97	1	69.03
5	6	0	0.00	0.0000	0	0	435.57	189721.22	0	22.54
6	7	0	0.00	0.0000	0	0	362.90	131696.41	0	32.97
7	8	0	0.00	0.0000	0	0	883.54	780642.93	0	22.48
8	9	1	599.30	6.9161	730	1	452.35	204620.52	1	17.98
9	10	1	271.18	6.0839	730	1	786.72	618928.36	1	38.91
10	11	0	0.00	0.0000	0	0	504.03	254046.24	1	28.85
11	12	0	0.00	0.0000	0	0	842.50	709806.25	0	49.41
12	13	0	0.00	0.0000	0	0	150.51	22653.26	1	41.91

The data is from Kumar and Petersen (2012), and refers to corporate clients.

Customer relationship management

Kumar and Petersen (2012) estimate a model to predict the response

$$Y = \begin{cases} 1 & \text{if the customer was acquired,} \\ 0 & \text{if the customer was not acquired,} \end{cases}$$

based on predictors such as the dollar spent on marketing efforts to acquire the prospect, and characteristics of the prospect's firm such as industry, revenue, and number of employees.

This is a binary classification problem.

Signet Bank

Around 1990, Richard Fairbanks and Nigel Morris realized that due to advances in data analysis they could do sophisticated predictive modeling for credit card customers.

Signet Bank, a small regional bank in Virginia took them on to model profitability and default probability.

They decided to invest in data: conducting experiments. Different credit card terms were offered *at random* to different customers.

The number of bad accounts soared and the bank lost money for a few years. But it viewed these losses as *investments in data*.

Eventually, Signet's credit card operation turned around and became highly profitable.

Notation

Notation

- We use upper case letters such as Y and X to denote random variables, regardless of dimension.
- Lower case letters denote observed values. For example, y denotes the realised value of the random variable Y.
- We use i to index the observations, j to index the inputs. For example, x_{ij} is the value of predictor j for observation i.
- We use the hat notation (e.g. $\widehat{\beta}$) for estimators and estimates.
- Vectors are in lower case bold letters. Matrices are in upper case bold letters.

Vector and matrix notation

Response vector:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Vector of observed values for the jth predictor (feature, input variable, attribute, covariate, regressor, independent variable):

$$\boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

Vector and matrix notation

We typically use n to denote the number of observations or cases, and we use p to denote the number of predictors.

Predictor matrix:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Statistical decision theory

Prediction

We seek a function $f(X)$ for predicting Y given the values of X.

We define prediction as follows:

1. Train a predictive function $\widehat{f}(\boldsymbol{x})$ using data $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$.

2. Upon observing a new input point $\boldsymbol{x}_0$, make the prediction $\widehat{f}(\boldsymbol{x}_0)$, i.e. the predictive function evaluated at $\boldsymbol{x}_0$.

How should we perform this learning task? How do we define our objective? How do we measure success in achieving this objective? To answer these questions, we turn to statistical decision theory.

Loss function

A **loss function** or **cost function** $L(y, f(\boldsymbol{x}))$ measures the cost making a prediction $f(\boldsymbol{x})$ when the truth is y. The most common loss function for regression is the **squared error loss**:

$$L(y, f(\boldsymbol{x})) = \Big(y - f(\boldsymbol{x})\Big)^2$$

For binary classification, a typical loss function is the **0-1 loss**:

$$L(y, f(\boldsymbol{x})) = \begin{cases} 1 \;\text{ if } y \neq f(\boldsymbol{x}) \\ 0 \;\text{ if } y = f(\boldsymbol{x}). \end{cases}$$

Expected loss

The idea of decision theory is to minimise the **expected loss** (or risk):

$$E\Big[L(Y, f(X))\Big].$$

Here, the expectation is over the joint probability distribution of X and Y, and f is treated as fixed.

The expected squared error loss, $E\left[(Y - f(X))^2\right]$, is minimized by the f that is defined as follows:

$$f(\boldsymbol{x}) = E(Y|X = \boldsymbol{x})$$

Concept: under the squared error loss, the optimal prediction of Y at any point $X = \boldsymbol{x}$ is the conditional mean $E(Y|X = \boldsymbol{x})$.

Statistical modelling

- Our regression problem reduces to the estimation of the conditional expectation function $E(Y|X = \boldsymbol{x})$. In order to learn this function, we need to introduce assumptions.

- Assumptions lead to statistical models.

- For example, the linear regression model assumes that $E(Y|X = \boldsymbol{x})$ is linear:

$$E(Y|X = \boldsymbol{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Additive error model

The **additive error model** is our basic general model for regression. It assumes that the relationship between Y and X is described as

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

where $f(.)$ is an unknown **regression function**, and ε is a random error with the expected value of zero.

Under this model,

$$E(Y|X = \boldsymbol{x}) = E(f(\boldsymbol{x}) + \varepsilon) = f(\boldsymbol{x}),$$

since $E(\varepsilon) = 0$.

Example: linear regression

In the special case of the linear regression model, we assume that

$$f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p,$$

leading to the model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon,$$

and predictions

$$\widehat{f}(\boldsymbol{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p,$$

where $\widehat{\beta}_j$ are least squares estimates of the regression coefficients.

Model Selection and Assessment

Statistical decision theory lays the foundation for:

- Choosing a learning method (model selection): estimating the performance of various learning methods and models in order to minimise the expected loss.

- Evaluating model performance (model assessment): estimating the expected loss of a chosen model.

Evaluating performance

Evaluating performance of a model

Model evaluation (or assessment) consists of estimating the expected loss of the trained model. To incorporate model assessment into our analysis, we split the dataset into two parts.

- **Training set**: for exploratory data analysis, model building, model estimation, etc.
- **Test set**: for model assessment.

Training and test data

- Because we are interested in estimating how well a model will predict future data, the test set should be kept in a “vault” and brought in strictly at the end of the analysis. The test set does not lead to model revisions.

- We generally allocate 50-80% of the data to the training sample.

- A higher proportion of training data leads to more accurate model estimation, but higher variance in estimating the expected loss.

Mean squared error

The choice of loss function leads to a measure of predictive accuracy. Suppose that we have observations y_i and predictions $\widehat{f}(\boldsymbol{x}_i)$ for an arbitrary sample, $i = 1, \ldots, m$. The **mean squared error** is:

$$\mathsf{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{f}(\boldsymbol{x}_i))^2$$

The test mean squared error is the MSE evaluated on the test set: $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^{m}$, for a given estimate $\widehat{f}$.

Test error

The test MSE is an estimate for the following quantity, which is known as **test** or **generalisation** error:

$$E\left[\left(Y - \widehat{f}(X)\right)^2\right].$$

Here, $\widehat{f}$ is treated as fixed, and the expectation is over the joint distribution of X and Y.

Note that the test error is also the expected loss for $\widehat{f}$, when we use the squared error loss and treat $\widehat{f}$ as fixed.

Mean squared error

Using the popular notation for predictions, $\widehat{y}_i = \widehat{f}(\boldsymbol{x}_i)$, we write

$$\mathsf{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y}_i)^2$$

The root mean-squared error and the prediction R^2 are derived from the MSE and are often a better way to report the results:

$$\mathsf{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \widehat{y}_i)^2}$$

$$\text{Prediction R}^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \overline{y})^2}$$

Overview of some key concepts

Some key concepts and themes

- Bias-Variance trade-off and model selection.
- Overfitting.
- Parametric vs non-parametric models.
- Accuracy vs interpretability.
- No-free lunch theorem.

Expected prediction error

- Focus on the squared error loss and consider the additive error model:

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

  where we assume that $\text{Var}(\varepsilon) = \sigma^2$.

- So far we have treated $\widehat{f}(\cdot)$ as given since our objective was to estimate the test error. Now, we discuss the fundamental problem of choosing a method to learn a predictive function $\widehat{f}(\cdot)$.

Expected prediction error

Think of $\widehat{f}$ as random, as it is constructed on a random training sample.

We define the **expected prediction error** for a new input point $X = \boldsymbol{x}_0$ as

$$E\left[\left(Y_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right]$$

where $Y_0 = f(\boldsymbol{x}_0) + \varepsilon_0$.

The expectation is over ε_0 and the training sample, i.e. over the sampling distribution of $\widehat{f}(\cdot)$.

Expected prediction error decomposition

In Tutorial 2 you will derive the following decomposition of the prediction error:

$$\begin{aligned} E\left[\left(Y_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] &= E\left[\left(f(\boldsymbol{x}_0) + \varepsilon_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] \\ &= \sigma^2 \qquad\qquad + \quad E\left[\left(f(\boldsymbol{x}_0) - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] \end{aligned}$$

Expected prediction error decomposition

$$\begin{aligned} E\left[\left(Y_0 - \widehat{f}(\boldsymbol{x}_0)\right)^2\right] &= \sigma^2 \qquad\qquad + \quad E\left[(f(\boldsymbol{x}_0) - \widehat{f}(\boldsymbol{x}_0))^2\right] \\ &= \text{Irreducible error} + \text{ Reducible error} \end{aligned}$$

- The first term is the variance of the response around its true mean $f(x_0)$. We cannot avoid this source of error, and it puts a bound on the accuracy of the prediction.

- In choosing a method, our concern is the reducible error: we want to minimise $E\left[(f(\boldsymbol{x}_0) - \widehat{f}(\boldsymbol{x}_0))^2\right]$.

The bias-variance trade-off

In Tutorial 2 you will derive the following relationship:

$$\begin{aligned}\text{Reducible Error} &= E\left[(f(\boldsymbol{x}_0) - \widehat{f}(\boldsymbol{x}_0))^2\right] \\ &= \left(E[\widehat{f}(\boldsymbol{x}_0)] - f(\boldsymbol{x}_0)\right)^2 + E\left[\left(\widehat{f}(\boldsymbol{x}_0) - E[\widehat{f}(\boldsymbol{x}_0)]\right)^2\right]\end{aligned}$$

The first term is the squared Bias in the estimation of $f(\boldsymbol{x}_0)$, we write it as $\text{Bias}^2\left(\widehat{f}(\boldsymbol{x}_0)\right)$

The second term is the Variance of the estimation, we write it as $\text{Var}\left(\widehat{f}(\boldsymbol{x}_0)\right)$

The bias-variance trade-off

$$E\left[\left(f(\boldsymbol{x}_0)-\widehat{f}(\boldsymbol{x}_0)\right)^2\right]=\text{Bias}^2\left(\widehat{f}(\boldsymbol{x}_0)\right)+\text{Var}\left(\widehat{f}(\boldsymbol{x}_0)\right)$$

- We would like our model to be flexible enough to be able to approximate complex relationships between X and Y.

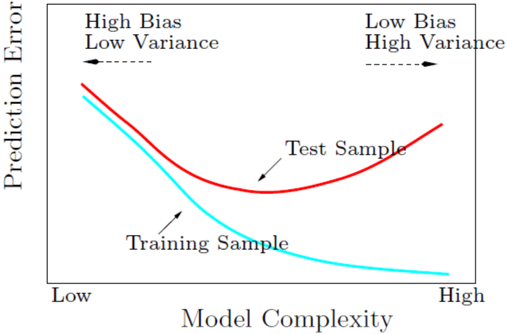
- Typically, the more complex we make the model, the better its approximation capabilities, which leads to lower bias.

- On the other hand, increasing model complexity leads to higher variance.

- Hence, we would like to find the optimal (problem specific) model complexity.

The bias-variance trade-off

Increasing model complexity will always reduce the training error, but there is an optimal level of complexity that minimises the test error.



Model selection

- **Model selection** is a process for choosing the right model among options of different complexity. It will be a fundamental part of our statistical learning toolkit.

- We conduct model selection on the training data.

Overfitting

- We say that there is **overfitting** when an estimated model is excessively flexible, incorporating minor variations in the training data that are likely to be noise rather than predictive patterns.

- A model that overfits has small training error, but may not generalize and perform poorly on new data.

- Avoiding being misled by overfitting is an important reason why we use a test set for assessment.

Illustration: predicting fuel economy

- This example uses data extracted from the fueleconomy.gov website run by the US government, which lists different estimates of fuel economy for passenger cars and trucks.
- For each vehicle in the dataset, we have information on various characteristics such as engine displacement and number of cylinders, along with laboratory measurements for the city and highway miles per gallon (MPG) of the car.
- Here we consider highway MPG as the response variable, and a single predictor: engine displacement.

Illustration: predicting fuel economy

A scatter plot reveals a nonlinear association between the two variables. We therefore need a model that is sufficiently flexible to capture this nonlinearity.

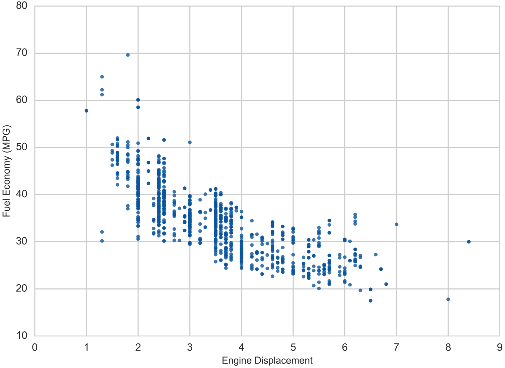
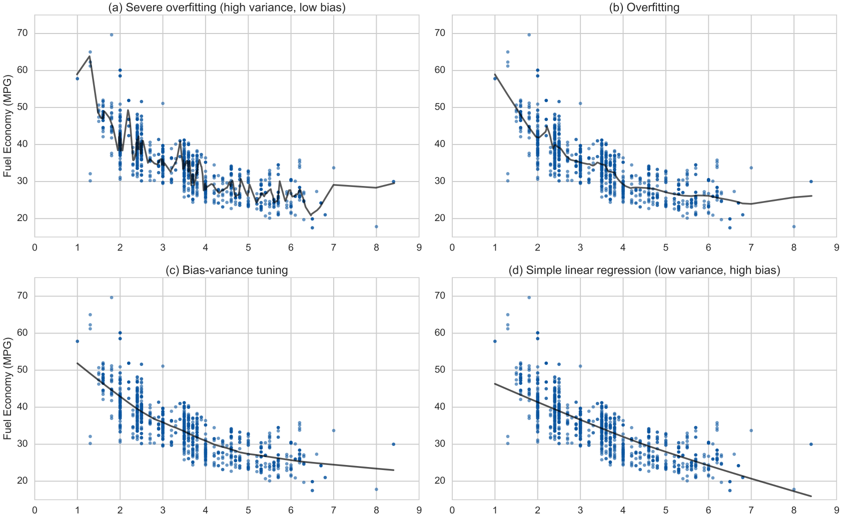


Illustration: predicting fuel economy



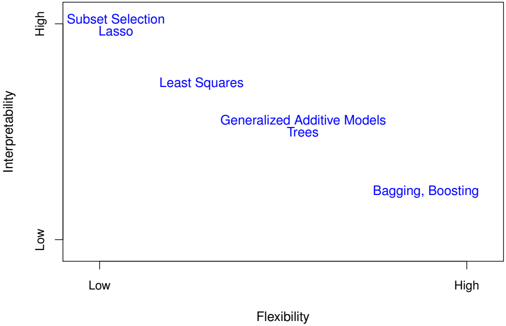
Parametric vs nonparametric models

There are many ways to define statistical models, but the most important distinction is the following:

- **Parametric models** assume that regression function f has a fixed number of parameters. Parametric models are faster to use, and more interpretable, but have the disadvantage of making stronger assumptions about the data.
- **Nonparametric models** do not impose the above assumption; the number of estimated parameters grows with the size of the training data. Nonparametric are more flexible, but have larger variance and can be computationally infeasible for large datasets.

Accuracy vs interpretability

Interpretability can be an important consideration in addition to predictive accuracy. Highly flexible, nonparametric methods, tend to be less interpretable than simpler methods.



No free lunch theorem

All models are wrong, but some are useful. – George Box

- The field of statistical learning proposes a large range of models and algorithms.
- However, there is no single model or approach that works optimally for all problems. This is sometimes called the **no free lunch theorem**.

Some review questions (1/2)

- What is the difference between supervised and unsupervised learning?
- What is a loss function?
- What do we learn from statistical decision theory for regression problems?
- How do we evaluate model performance with data?

Some review questions (2/2)

- What is the bias-variance trade-off and why is it important for predictive modelling?
- What is model selection? How is it different from model evaluation (assessment)?
- What is overfitting?
- What is the difference between parametric and nonparametric models? What are the advantages and disadvantages of each approach?