ETC3250: Introduction
Semester 1, 2019

Professor Di Cook
Econometrics and Business Statistics
Monash University
Week 1 (b)

Outline

Better understand or make predictions about a certain phenomenon under study

Construct a model of that phenomenon by finding relations between several variables

If phenomenon is complex or depends on a large number of variables, an analytical solution might not be available

However, we can collect data and learn a model that approximates the true underlying phenomenon
Outline

- Introduction
  - Learning from data

\[ D = \{ (x_i, y_i) \}_{i=1}^{n} \text{ where } x_i = (x_{i1}, \ldots, x_{id})^T \]

**Statistical learning** provides a framework for constructing models from the data.

---

Outline

- Introduction
  - Learning from data
  - Different learning problems

- Supervised learning, \( y \) **available** for all \( x \).
  - Regression (or prediction)
  - Classification

- Unsupervised learning, \( y \) **unavailable** for all \( x \).
- Semi-supervised learning, \( y \) available only for few \( x \).
- Other types of learning: reinforcement learning, online learning, active learning, etc.

**Being able to identify which is the type of learning problem you have is important in practice.**
\[ \mathcal{D} = \{(y_i, x_i)\}_{i=1}^n \]

where \( (y_i, x_i) \sim P_{\mathbf{y}, \mathbf{x}} = P(\mathbf{y}|\mathbf{x}) \) means that these arise from some probability distribution. \( \sim \) means distributed as, arise from. Typically, we only are interested in \( P(\mathbf{y}|\mathbf{x}) \), the distribution of \( \mathbf{y} \) conditional on \( \mathbf{x} \).

---

**Outline**

- Introduction
- Supervised learning

\[ y = (y_1, \ldots, y_q); \text{ response (output) (could be multivariate, } q = 1 \text{ for us)} \]

\[ x = (x_1, \ldots, x_p); \text{ set of } p \text{ predictors (input)} \]

We seek a function \( h(x) \) for predicting \( y \) given values of the input \( x \). This function is computed using \( \mathcal{D} \).
We are interested in minimizing the expected out-of-sample prediction error:

$$\text{Err}_{\text{test}}(h) = E[L(h(Y), h(X))]$$

where $L(y, \hat{y})$ is a non-negative real-valued loss function, such as $L(y, \hat{y}) = (y - \hat{y})^2$ and $L(y, \hat{y}) = I(y \neq \hat{y})$.

The goal is that the predictions from the model are accurate for future samples.

---

We often assume that our data arose from a statistical model

$$Y = f(X) + \epsilon, \text{ for some function } f, \text{ where } \epsilon \text{ is the random error term with } E[\epsilon] = 0 \text{ and is independent of } X.$$  

The additive error model is a useful approximation to the truth

$$f(x) = E[Y | X = x]$$

Not a deterministic relationship: $Y \neq f(X)$
Blue curve is $f(x)$, the true functional relationship.
Outline

- Introduction
- Supervised learning
  - Regression
  - Why?
  - Estimation

**Prediction:**
- $\hat{y} = f(x)$ for a new observation $x$.

**Inference (or explanation):**
- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?

---

**Linear model:**
$f(\text{education, seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \text{education} + \hat{\beta}_2 \times \text{seniority}$

Why would we ever choose to use a **more restrictive method** instead of a **very flexible approach**?

|Chapter2/2.3.pdf, 2.4.pdf|
Outline

- Introduction
- Supervised learning
  - Regression
  - Why?
  - Estimation

- Parametric methods
  - Assumption about the form of \( f \), e.g., linear
  - The problem of estimating \( f \) reduces to estimating a set of parameters
  - Usually a good starting point for many learning problems
  - Poor performance if linearity assumption is wrong

- Non-parametric methods
  - No explicit assumptions about the form of \( f \), e.g., nearest neighbours: \( Y(x) = \frac{1}{n} \sum_{i=1}^{n} y_i \)
  - High flexibility: it can potentially fit a range of shapes
  - A large number of observations is required to estimate \( f \) with good accuracy
Suppose we have a regression model $y = f(x) + \epsilon$. Estimate $f$ from some training data, $\mathcal{T}_r = \{(x_i, y_i)\}_{i=1}^{n_r}$.

One common measure of accuracy is: 

**Training Mean Squared Error**

$$MSE_{\mathcal{T}_r} = \frac{1}{n_r} \sum_{i=1}^{n_r} (y_i - f(x_i))^2$$

Measure real accuracy using test data $\mathcal{T}_e = \{(x_i, y_i)\}_{i=1}^{n_e}$.

**Test Mean Squared Error**

$$MSE_{\mathcal{T}_e} = \frac{1}{n_e} \sum_{i=1}^{n_e} (y_i - f(x_i))^2$$
Outline

- Introduction
- Supervised learning
- Assessing model accuracy
  - Regression
  - Training vs Test MSEs

In general, the more flexible a method is, the lower its training MSE will be. i.e. it will “fit” the training data very well.

However, the test MSE may be higher for a more flexible method than for a simple approach like linear regression.

Flexibility also makes interpretation more difficult. There is a trade-off between flexibility and model interpretability.

Outline

- Introduction
- Supervised learning
  - Regression
  - Why?
  - Estimation
  - Methods
  - Interpretability vs flexibility

Simplistic overview of methods on the flexibility vs interpretability scale. Interpretability is when it is clear how the explanatory variable is related to the response, e.g. linear model. Poor interpretability is often called a "black box."
Outline

- Introduction
- Supervised learning
- Assessing model accuracy
  - Regression
  - Training vs Test MSEs
  - Example

true curve
linear regression

Smoothing splines

Training MSE
Test MSE

Dashed: Minimum test MSE
Outline

Introduction
Supervised learning
Assessing model accuracy
  - Regression
    - Training vs Test MSEs
    - Example

true curve
linear regression
Smoothing splines

Training MSE
Test MSE
Dashed: Minimum test MSE

There are two competing forces that govern the choice of learning method: bias and variance.

Bias is the error that is introduced by modeling a complicated problem by a simpler problem.

- For example, linear regression assumes a linear relationship when few real relationships are exactly linear.
- In general, the more flexible a method is, the less bias it will have.

This site has a lovely explanation, if you don’t like mine.
There are two competing forces that govern the choice of learning method: bias and variance.

Variance refers to how much your estimate would change if you had different training data.

In general, the more flexible a method is, the more variance it has.

The size of the training data has an impact on the variance.

---

**MSE decomposition**

If $Y = f(x) + e$ and $f(x) = E[Y | X = x]$, then the expected test MSE for a new $Y$ at $x_0$ will be equal to

$$E[(Y - f(x_0))^2] = \text{Bias}^2 + \text{Var}(f(x_0)) + \text{Var}(e)$$

**Test MSE = Bias + Variance + Irreducible variance**

1. The expectation averages over the variability of $Y$ as well as the variability in the training data.
2. As the flexibility of $f$ increases, its variance increases and its bias decreases.
3. Choosing the flexibility based on average test MSE amounts to a bias-variance trade-off
squared bias, variance, \( \text{Var}(e) \) (dashed line), and test MSE for the three data sets shown earlier. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

(Chapter 2/2.12 pdf)
Here the response variable $y$ is qualitative.

- e.g., email is one of $c = \{spam, ham\}$
- e.g., voters are one of $c = \{Liberal, Labor, Green, National, Other\}$

Our goals are:

1. Build a classifier $f(y)$ that assigns a class label from $c = \{c_1, \ldots, c_K\}$ to a future unlabeled observation $x$.
2. Such a classifier will divide the input space into regions $R_k$ called decision regions, one for each class, such that all points in $R_k$ are assigned to class $c_k$.
3. Assess the uncertainty in each classification (i.e., the probability of misclassification).
4. Understand the roles of the different predictors among $X = (x_1, x_2, \ldots, x_p)$.

Recall that we want to minimize the expected prediction error

$$\mathbb{E}_{y,x} [L(Y, f(X))]$$

where $L(y, \hat{y})$ is a non-negative real-valued loss function.

In classification, the output $y$ is a categorical variable, and our loss function can be represented by a $K \times K$ matrix $L$, where $K = \text{card}(c)$. $L(k, \hat{k})$ is the cost of classifying $c_k$ as $c_{\hat{k}}$. With zero-one loss, i.e., $L(y, \hat{y}) = I(y \neq \hat{y})$, the cost is equal for mistakes between any class.
Outline

- Introduction
- Supervised learning
- Assessing model accuracy
  - Classification
  - Optimal classifier
  - Error

**Compute \( \hat{c} \) from some training data, \( \mathcal{T}_t = (x_t, y_t) \).**

In place of MSE, we now use the error rate (fraction of misclassifications).

**Training Error Rate**

\[
\text{Error rate}_t = \frac{1}{\| \mathcal{T}_t \|} \sum_{(x_t, y_t) \in \mathcal{T}_t} I(y_t \neq \hat{c}(x_t))
\]

Measure **real accuracy using test data** \( \mathcal{T}_e = (x_e, y_e) \).

**Test Error Rate**

\[
\text{Error rate}_e = \frac{1}{\| \mathcal{T}_e \|} \sum_{(x_e, y_e) \in \mathcal{T}_e} I(y_e \neq \hat{c}(x_e))
\]

---

Outline

- Introduction
- Supervised learning
- Assessing model accuracy
  - Classification
  - Error
  - Bayes classifier

Let \( c = \{c_1, \ldots, c_K \} \), and let

\[
p_k(x) = P(Y = c_k | X = x), \quad k = 1, 2, \ldots, K.
\]

These are the **conditional class probabilities** at \( x \).

Then the **Bayes classifier** at \( x \) is

\[
\hat{c}(x) = C_k \quad \text{if } p_k(x) = \max(p_1(x), p_2(x), \ldots, p_K(x))
\]

**This gives the minimum average test error rate.**

---

28 / 38

---

29 / 38
The Bayes error rate is the lowest possible error rate that could be achieved if we knew exactly the true probability distribution of the data. It is analogous to the irreducible error in regression. On test data, no classifier can get lower error rates than the Bayes error rate. In reality, the Bayes error rate is not known exactly.
One of the simplest classifiers. Given a test observation $x_t$, find the $K$ nearest points to $x_t$ in the training data: $N_t$.

Estimate conditional probabilities:

$$P(Y = C_j | X = x_t) = \frac{1}{K} \sum_{i=1}^{K} I(y_i = C_j).$$

Classify $x_t$ to class with largest probability.
Outline

- Introduction
- Supervised learning
- Assessing model accuracy
  - Classification
  - Error
  - Bayes classifier
  - kNN

KNN: K=1

KNN: K=100

KNN: K=10

(Chapter 2/2.16.pdf)
Outline

- Introduction
- Supervised learning
- Assessing model accuracy
  - Classification
  - Error
  - Bayes classifier
  - kNN

A fundamental picture
Made by a human with a computer

Slides at https://monba.dicook.org.
Code and data at https://github.com/dicook/Business_Analytics.

Created using R Markdown with flair by xaringan, and kunoichi (female ninja) style.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.