Predictive Analytics

Week 1: Introduction to Predictive Modelling

Semester 2, 2018

Discipline of Business Analytics, The University of Sydney Business School

Week 1: Introduction to Predictive Modelling

1. Content structure
2. Introduction
3. Business examples and data
4. Notation
5. Statistical decision theory
6. Evaluating model performance
7. Key concepts and themes

Content structure

QBUS2820 content structure

1. **Statistical and Machine Learning foundations and applications.**

2. **Advanced regression methods.**

3. **Classification methods.**

4. **Time series forecasting.**

Content structure

1. Statistical Machine Learning foundations and applications: key concepts in predictive modelling, statistical thinking. K-nearest neighbours, model evaluation, model selection, and model inference, etc.

2. Regression: subset selection, ridge regression, LASSO, principal components regression, etc.

3. Classification: key concepts, evaluating classification models, logistic regression, regularised logistic regression, linear and quadratic discriminant analysis, etc.

4. Forecasting: key concepts, time series, exponential smoothing and ARIMA models, etc.

Learning outcomes

By successfully completing this unit, you are expected to:

1. Understand the conceptual and theoretical foundations of predictive modelling.
2. Develop an in-depth knowledge of basic methods for regression, classification and forecasting 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 (including tasks such as data manipulation and visualisation).
5. Effectively communicate your results to guide decision making.

Comments

- This unit is designed as training for real-world predictive analytics, which requires a range of skills.

- Practical work in this area involves more than knowing the methods in the lectures: professionals typically spend a substantial amount of time on tasks such as data management, exploratory data analysis, feature engineering, and implementing methods.

- All of this generally done through coding. Therefore, Python is your bridge between knowledge and practice.

- For these reasons, please note that this unit requires independent work and higher than average workload (within the university guidelines).

Introduction

Predictive modelling is a set of methods for detecting patterns in data and using these patterns for predicting future data and informing decision making. In this unit, we will draw on methods from the fields of statistics, econometrics, and machine learning.

Introduction

Two trends bring predictive modelling 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.

- Advancing computing power (realising **Moore's law**) increases the scope for exploring complex patterns in data.

Types of prediction

Different types of data lead to different types of prediction problems:

- In **cross sectional prediction**, we work with data collected by observing subjects (such as individuals, firms, assets, etc). Our objective is to predict the value of a response variable for a new subject.

- In **forecasting**, we want to predict the value of a response variable at specific point in the future, based on past and current information. Forecasting can be based on time series data for the response variable only.

Types of learning

- Supervised learning
- Unsupervised learning

Supervised learning

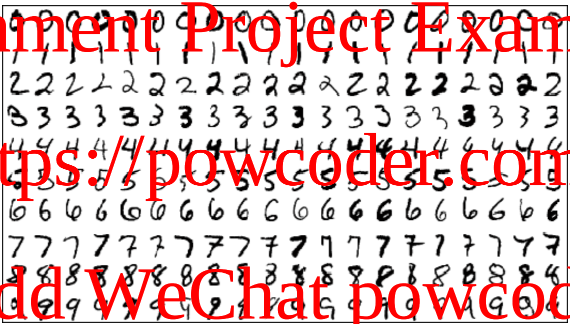
In the context of statistical learning, **supervised learning** is the task of learning a function to predict an output variable Y based on observed input variables $x_1, \ldots, x_p$. We develop methods that learn this function based on labelled data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, which we call the training data.

Supervised learning

In supervised learning, the output or **response variable** can be of any type. We will study methods that address two main classes of supervised learning problems:

- In **regression**, the response is a quantitative scalar (such as the income of a worker).
- In **classification**, the response is **nominal** or **categorical** variable $Y \in \{1, \ldots, C\}$, where C is the number of classes. When $C = 2$, this is called binary classification; if $C > 2$, this is called multiclass classification.

Example: handwritten digit recognition



A view of the MNIST dataset.

Unsupervised learning

Unsupervised Learning: No distinction is made between Y and X. "Unlabelled" data is used to uncover hidden patterns, clusters, relationships or distribution

- E.g. Principle Component Analysis: Aiming to find the key factors determining data patterns
- Goal: Hypothesis generation, then to be tested in supervised learning

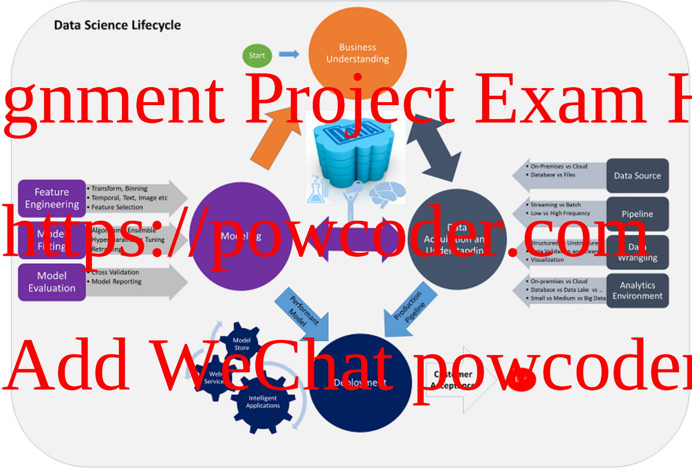
Learner: A learner is a (mathematical) model for learning, e.g. estimated a regression model based on a training data set.

Data science

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) link their analysis to actionable results. A common objective is to create **data products**.

The data science process: a real-world perspective



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

Data analysis process in this unit

1. Problem formulation.
2. Data collection and preparation.
3. Exploratory data analysis (EDA).
4. Model building, estimation, and selection.
5. Model evaluation.
6. Communicate results.

Assignment Project Exam Help

Business examples and data

https://powcoder.com

Add WeChat powcoder

Examples

- Credit card fraud detection: collect data from multiple sources to learn typical customer behaviour, then use this model to detect suspicious transactions for further investigation.
- Customer risk analysis: instead of denying sales (say, auto loans, credit cards, and insurance policies) to higher risk customers, it is usually a better strategy to price risk accordingly using available data.
- Advertising: making online ads more relevant to users by predicting click-through rates.

Zillow Kaggle competition

- Kaggle is a crowdsourcing platform that allows organisations to post data prediction problem to be solved by public competition.

- Zillow's Home Value Prediction is a current competition (with a 1.2 million dollar cash prize) that invites participants to make predictions about the future sale prices of homes (a regression problem).

- In this competition, the goal is 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 that involve 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 may be part of a customer-centric (as opposed to brand-centric) business strategy, which focuses on customer satisfaction and loyalty towards 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 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	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]
2	3	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	[illegible]
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	[illegible]	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	[illegible]	0	32.97
7	[illegible]	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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	[illegible]	0	0	842.50	709806.25	0	49.41
12	[illegible]	0	[illegible]	[illegible]	[illegible]	0	[illegible]	[illegible]	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 employees.

This is a binary classification problem.

Notation

Study tips

- Always start by making sure that you understand the notation and definitions. Focus first on meaning, then connections.

- If there is a learning challenge, is the root of the problem in understanding notation, concepts, reasoning, or algebra?

- When reading an equation, you should be able to identify parameters and constants, distinguish between random variables and observed values, and distinguish between scalars, vectors, and matrices.

- When there is an expectation or variance operator, what distribution is it over? That is, what random variables do they refer to?

Notation

- We use upper case letters such as Y 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, y_i is the observed response for sample i, while x_{ij} is the value of predictor j for observation i.

- We use the hat notation (e.g. $\hat{\beta}$) for estimators and estimates. The notation may not distinguish between the two (refer to context).

- Vectors are in lower case bold letters. Matrices are in upper case bold letters.

Vector and matrix notation

Response vector:

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

Review the provided materials of liner algebra.

Vector and matrix notation

Vector of predictor (features, attributes, covariates, regressors, independent variables) values for observation i:

$$\boldsymbol{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Vector of observed values for predictor j:

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

Vector and matrix notation

Design matrix:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}$$

Statistical decision theory

Prediction

We define prediction as follows:

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

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

How should we perform this prediction task? How do we define our objective? How do we measure success in achieving this objective? To answer these questions, we turn to **decision theory**. We mostly focus on regression problems for simplicity.

Loss function

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

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

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

$$L(y, \widehat{y}) = \begin{cases} 1 & \text{if } y \neq \widehat{y} \\ 0 & \text{if } y = \widehat{y}, \end{cases}$$

where $\widehat{y}$ is the prediction.

Expected loss

Let Y and X have a joint probability distribution $P(X, Y)$. The idea of decision theory is that we take the action that minimises our **expected loss** or **risk**:

$$R(f) = E\left[L(Y, f(X))\right],$$

where the expectation is over $P(X, Y)$. Here, the risk is for a given function $f(\cdot)$.

We can use the **law of iterated expectations** to rewrite the expected loss as

$$R(f) = E\left[E\left(Y - f(X)\right)^2 | X\right].$$

Optimal prediction

The optimal action is to choose the prediction function $\delta(.)$ that minimises the expected loss. This is equivalent to minimising the expected loss at every input point $\boldsymbol{x}$:

$$\delta(\boldsymbol{x}) = \underset{f(.)}{\text{argmin}}\ E(L(Y, f(\boldsymbol{x}))|X = \boldsymbol{x})$$

The solution for the squared loss (see module notes) is the conditional expectation:

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

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:

$$E(Y|X = \boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$$

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 mean zero ($E(\varepsilon) = 0$).

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 $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ is the vector of least squares estimates of the model parameters.

Statistical decision theory

Our discussion of statistical decision theory lays the foundation for the rest of our discussion.

- Evaluating model performance: estimating the expected loss of a trained model.

- Choosing a learning method: finding and estimating an appropriate model such that we minimise our expected loss.

Evaluating model performance

Evaluating model performance

Model evaluation consists of estimating the expected loss of a trained model. To incorporate model assessment into our analysis, we split the dataset into three parts.

- **Training set**: for exploratory data analysis, model building, model estimation, model selection, etc.
- **Validation set**: for appropriate model selection.
- **Test set**: for model evaluation.

Training, validation and test data

- Because we are interested on the 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.

- The split of the data into the training, validation and test sets is often random, but sometimes there are reasons to consider alternative schemes.

Evaluating test performance

Suppose that we have test observations $\{(\widetilde{y}_i, \widetilde{\boldsymbol{x}}_i)\}_{i=1}^M$ and corresponding predictions $\widehat{f}(\widetilde{\boldsymbol{x}_i})$ for $i = 1, \ldots, M$. We evaluate model performance by computing the **empirical risk** for the test set:

$$\widehat{R}_{\text{test}} = \frac{1}{M} \sum_{i=1}^{M} L\left(\widetilde{y}_i, \widehat{f}(\widetilde{\boldsymbol{x}_i})\right)$$

Below, we drop the specific notation for test observations for simplicity.

Mean squared error

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

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

The test mean squared error is the MSE evaluated for the test set.

Mean squared error

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

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

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

Mean absolute error

Another common measure of performance is the **mean absolute error** (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i|$$

- Implicit in the use of the MAE is the absolute error loss function. The absolute error setting is less mathematically tractable, which is one of the reasons why focus on the squared error loss.

- In this case the optimal prediction is the conditional median, not the mean.

The **test** or **generalisation** error is the expected loss for the model estimated with the training data $\mathcal{D}$. We define it as

$$\text{Err} = E\left[L\left(Y, \widehat{f}(X)\right) \middle| \mathcal{D}\right],$$

where the expectation is over $P(X, Y)$.

Concept: the test MSE estimates the test error (under the squared error loss).

Standard error

As always, you should report a measure of sample uncertainty for every important estimate in your analysis. The test MSE is a sample average, so obtaining a standard error is straightforward. The formula for a general sample is:

$$\mathrm{SE(MSE)} = \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}\frac{\left(\left(y_i - \widehat{f}(\boldsymbol{x}_i)\right)^2 - \mathrm{MSE}\right)^2}{n-1}}$$

Inference for the test errors is possible, but we do not pursue this here.

Key concepts and themes

Key concepts and themes

- Underfitting and Overfitting.

- Parametric vs non-parametric models.
- No-free lunch theorem.

- Accuracy vs interpretability.

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.
- An overfit model has small training errors, but may predict poorly. In essence, it has memorised the training set.
- Not being misled by overfitting is an important reason why we use a test set.
- We will present more details about bias variance decomposition later.

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.
- We here consider the unadjusted highway MPG for 2010 cars 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.

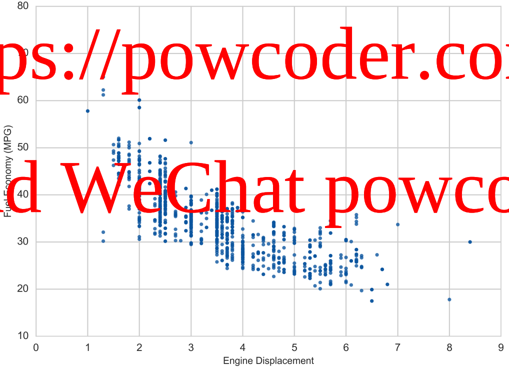
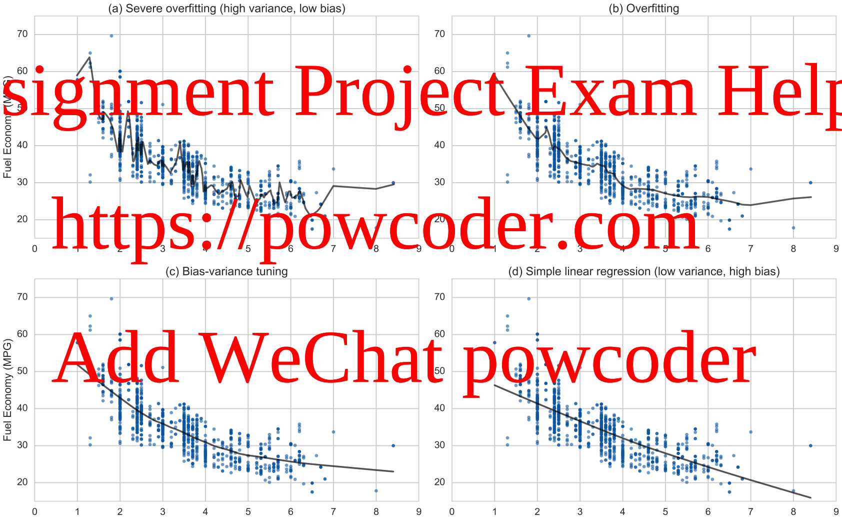


Illustration: predicting fuel economy



Parametric vs nonparametric models

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

- A **parametric model** 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.

- In a **nonparametric model**, the number of 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. An example is the K-nearest neighbours method, which we will study in the next module.

No free lunch theorem

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

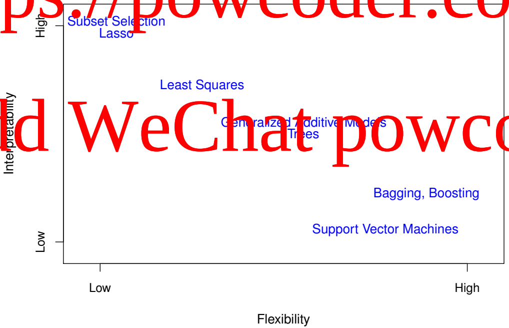
- The field of machine learning proposes a large range of models and algorithms to solve supervised and unsupervised learning problems.

- However, there is no single model or approach that works optimally for all problems. This is sometimes called the **no free luch theorem**.

- Therefore, applied statistical learning requires awareness of speed-accuracy-complexity trade-offs and data-driven consideration of different approaches for every problem.

Accuracy vs interpretability

Particularly in data mining, interpretability is an important consideration in addition to predictive accuracy. Highly flexible, nonparametric methods, tend to be less interpretable than simpler methods.



Study guide

- Recall three important concepts from these slides, and explain them in your own words.

- Use the review questions in the next slide to self-test on key concepts.
- Study the mathematical details in the module notes.

- Study (or revise) Chapters 1 and 2 of ISL. Reader Chapter 3 before the next module.

Review questions (1/2)

- What is predictive modelling?

- What is the difference between cross-sectional prediction and forecasting?

- What is supervised 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?

Review questions (2/2)

- What is the difference between the generalisation error and the expected prediction error?
- 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?
- What is overfitting?
- What is the difference between parametric and nonparametric models? What are the advantages and disadvantages of each approach?