Statistical learning - Introduction

Rita Almeida

PhD course 3045, VT 2018

Statistical modeling / Machine learning

- What are the differences?
  - well...

Statistics	Machine Learning
Estimation / Fitting	Learning
Data point	Example
Regression / Classification	Supervised Learning
Clustering	Unsupervised Learning
Parameters	Weights
Covariate	Feature
Response	Label
Test performance	Generalization
Inference	Prediction
Small data sets	Large data sets
Light computation	Heavy computation
Large grant = $50,000	Large grant = $1,000,000
Old / Boring	New / Cool / Vibrant
R	Python

Adapted from Larry Wasserman and Rob Tibshirani

Statistical learning

Statistical modeling: emphasizes statistical inference in low dimensional problems.

Machine learning: emphasizes prediction in high dimensional problems.

Statistical learning: set of tools for modeling and understanding data.

Supervised and unsupervised learning

Supervised learning: Finding a model that relates a response variable (label) to a set of predictor variables.

- Inference
- Prediction

- Regression
- Classification

- Parametric
- Non-parametric

Unsupervised learning: Finding a hidden structure in a set of data without a response variable (unlabeled).

- Cluster analysis
- Principal Component Analysis (PCA)

Supervised learning

- Types of problems:
  - Inference and prediction
  - Regression and classification
  - Parametric and non-parametric
- Assessing model accuracy
  - Measuring quality of fit
  - Bias-variance trade-off
  - Cross-validation

Supervised learning

- Types of problems:
  - Inference and prediction
  - Regression and classification
  - Parametric and non-parametric

- Assessing model accuracy
  - Measuring quality of fit
  - Bias-variance trade-off
  - Cross-validation

Inference and Prediction

In many problems one wants to estimate the relation between:

- input variables

  explanatory variables
  independent variables
  features
  covariates

- output variable

  outcome
  response
  dependent variable
  label

$$Y = f(X) + \epsilon \qquad X = (X_1, X_2, \ldots, X_p)$$

Why estimating f?

- Inference
- Prediction

Inference

Aim: understanding how Y is influenced by the Xs.

- Which explanatory variables are important? Which are associated with the outcome?
- How is the relation between each explanatory variable and the outcome?
  - Positive or negative association
  - Linear or other

Example:

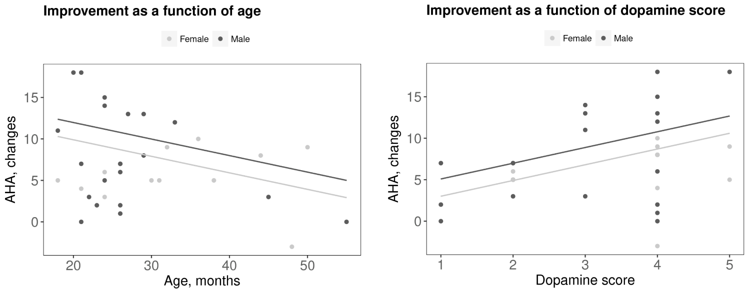


Figure adapted from Diaz-Heijtz et al. 2018

Prediction

Aim: predict value of Y given a set of Xs.

- Typically the exact form of f does not matter as long as we get accurate predictions.
- We want to estimate f in a way that minimizes the reducible error.

The accuracy of the prediction $\hat{Y}$ of Y depends on:

- reducible error: error in the estimate $\hat{f}$ of f
- irreducible error: ϵ - there are unmeasured factors that influence the outcome

Supervised learning

- Types of problems:
  - Inference and prediction
  - Regression and classification
  - Parametric and non-parametric
- Assessing model accuracy
  - Measuring quality of fit
  - Bias-variance trade-off
  - Cross-validation

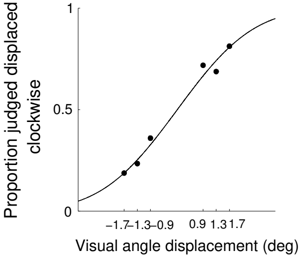
Regression and Classification

Types of variables:

- quantitative
  - examples: age, height, income
- qualitative
  - examples: language spoken, existence of a diagnose

Outcome variables

- quantitative - regression
- qualitative - classification
- But things are not always so clear...



Explanatory variables can be quantitative or qualitative for both types of problems.

Supervised learning

- Types of problems:
  - Inference and prediction
  - Regression and classification
  - Parametric and non-parametric
- Assessing model accuracy
  - Measuring quality of fit
  - Bias-variance trade-off
  - Cross-validation

How to estimate f: parametric approaches

Parametric approaches:

1. assume a parametrized form for f
2. estimate the parameters

- Reduces and simplifies the problem of estimating f!

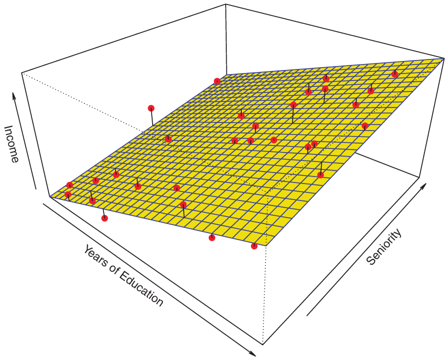


Figure adapted from James, Witten, Hastie, Tibshirani

For example a linear model using least squares for estimation:

1. assume $Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$
2. estimate the βs

How to estimate f: non-parametric approaches

Non-parametric approaches:

- Do not make explicit assumptions about f, but impose smoothness constrains.

  Advantage: a potential wrong shape of f is not pre-defined.

- The problem of estimating f is not reduced.

  Disadvantage: a lot more data is required to estimate f.

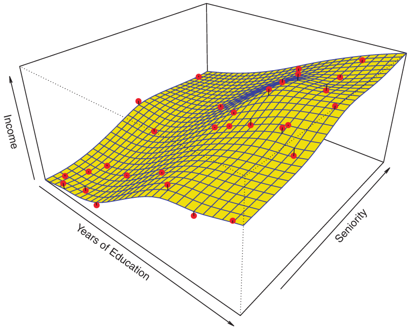


Figure adapted from James, Witten, Hastie, Tibshirani

Model interpretability versus prediction accuracy

- What approach for estimating f gives better interpretability?
- What approach for estimating f gives better prediction accuracy?

- For what problems does one favor interpretability?
- For what problems does one favor accuracy?

Model interpretability versus prediction accuracy

- More interpretability:
  - Parametric approaches - more inflexible models

    Least squares linear model

    Lasso (least absolute shrinkage and selection operator)
  - Inference problems

- Better prediction accuracy:
  - Non-parametric approaches - more flexible models

    Support-vector classification
  - Prediction problems

Supervised learning

- Types of problems:
  - Inference and prediction
  - Regression and classification
  - Parametric and non-parametric
- Assessing model accuracy
  - Measuring quality of fit
  - Bias-variance trade-off
  - Cross-validation

Quality of fit - regression

There is no single best method for all data sets and questions.

Inference

Mean squared error is commonly used:

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

- MSE (training MSE) should be small.

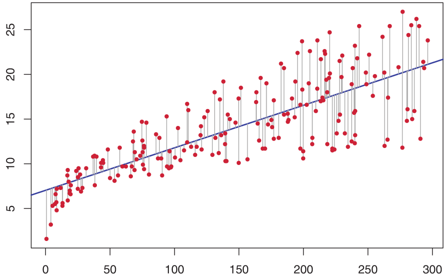


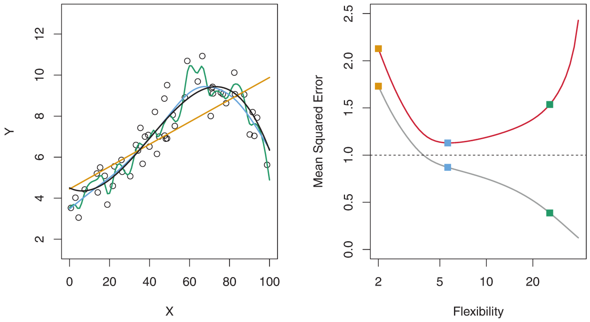
Figure adapted from James, Witten, Hastie, Tibshirani

Prediction:

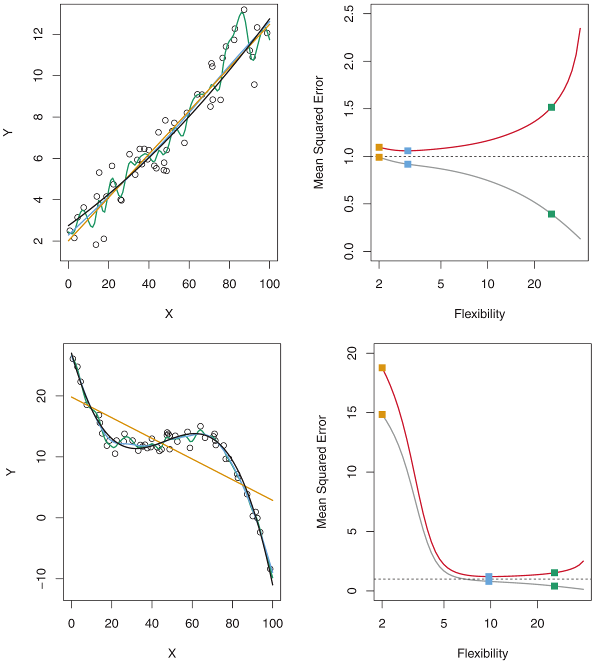
- For a new observation x_0, $\hat{f}(x_0)$ is close to y_0.
- MSE on test data (test MSE) should be small.
- Average $(y_0 - \hat{f}(x_0))^2$ should be small.

Overfitting

Minimizing training MSE does not guarantee minimization of test MSE



Figures adapted from James, Witten, Hastie, Tibshirani



Y
X
0
20
40
60
80
100
2
4
6
8
10
12
Mean Squared Error
Flexibility
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Y
X
−10
0
10
20
0
20
40
60
80
100
Mean Squared Error
Flexibility
0
5
10
15
20
2
5
10
20

Overfitting

Linear model, fitting polynomials.

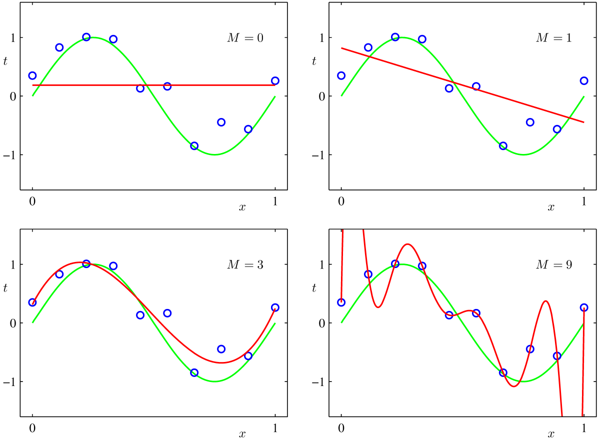


Figure adapted from Bishop

Supervised learning

- Types of problems:
  - Inference and prediction
  - Regression and classification
  - Parametric and non-parametric

- Assessing model accuracy
  - Measuring quality of fit
  - Bias-variance trade-off
  - Cross-validation

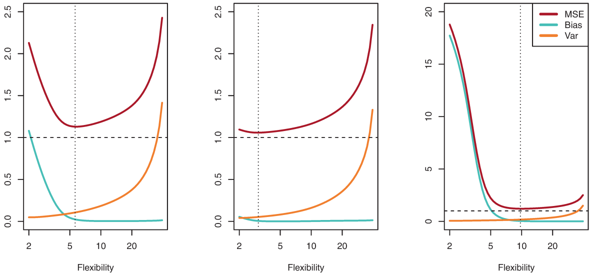
Bias-variance trade off

Variance: amount by which $\hat{f}$ changes when estimated based on other data sets

- flexible models $\rightarrow$ more variance

Bias: error introduced by using a given model

- flexible models $\rightarrow$ less bias



Figures adapted from James, Witten, Hastie, Tibshirani

Quality of fit - classification

Inference

Error rate is commonly used:

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

$\hat{y}_i$ is the label predicted by $\hat{f}$

$$I(y_i \neq \hat{y}_i) \begin{cases} 1 & \text{if } y_i \neq \hat{y}_i \\ 0 & \text{if } y_i = \hat{y}_i \end{cases}$$

- Training error rate should be small.

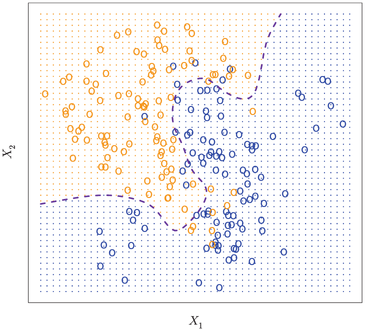


Figure adapted from James, Witten, Hastie, Tibshirani

Prediction:

- For a new observation x_0, $\hat{y}_0$ is the correct label.
- Error rate on test data (test error rate) should be small.
- $\sum I(y_i \neq \hat{y}_i)$ should be small.

Overfitting and bias-variance trade off

Overfitting and bias-variance trade-off also apply to classification problems.

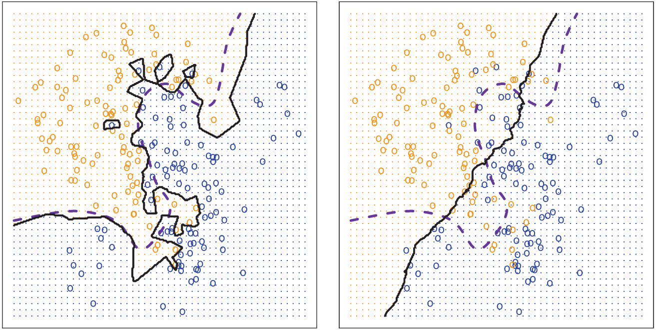


Figure adapted from James, Witten, Hastie, Tibshirani

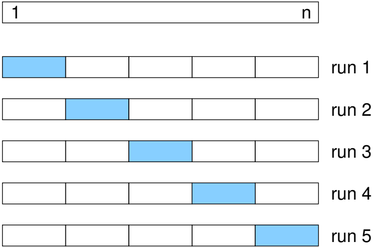
Supervised learning

- Types of problems:
  - Inference and prediction
  - Regression and classification
  - Parametric and non-parametric
- Assessing model accuracy
  - Measuring quality of fit
  - Bias-variance trade-off
  - Cross-validation

Cross-validation

How to compute the test MSE / error rate if there is no test set?

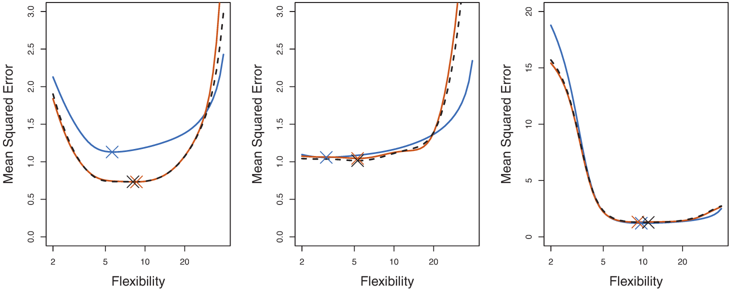
- Divide the data into training and test data.
  - If the data set is small - too few observations to train and too few observation to estimate the quality of fit.
- k-fold cross-validation
  - Divide the data into k subsets.
  - For each subset k train the model on the rest of the data and test on k.
  - Average the k MSEs or error rates.
  - If $k = n$ one observation is left out: Leave-one-out cross-validation



Cross-validation

Aims:

- Estimate how accurate is a given model.
- Identify what statistical learning method is best for a given data.



orange: 10-fold; dashed: leave-one-out

Figure adapted from James, Witten, Hastie, Tibshirani

Bias-variance trade-off for k-fold cross-validation

What k to choose for cross-validation?

- Dividing the data in 2 $\rightarrow$ half the data is used to train $\rightarrow$ overestimation of the test error (bias)
- Large k $\rightarrow$ data used for different runs very similar $\rightarrow$ high variance of the test error estimate
- $k = 5$ or 10 are commonly used.

Overview of the rest of the day

- Regression
  - Shrinkage Methods: Ridge regression, Lasso
  - Non-linear regression: polynomial regression, splines
- Classification
  - Nearest-neighbors
  - Logistic and multinomial regression
  - Discriminant analysis
  - Support vector classification
- Non-supervised
  - Principal component analysis
  - Cluster analysis

References

- G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning , with applications in R*, Springer texts in statistics
- C. Bishop, *Pattern Recognition and Machine Learning*, Springer