Analysis Methodologies: essential machine learning and overview

Introduction to Big Data

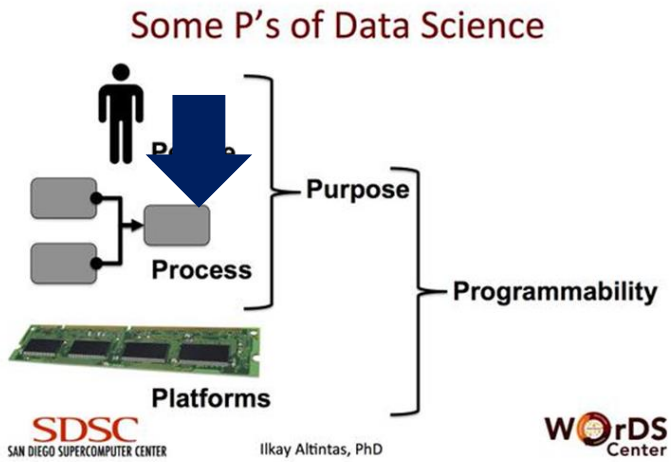
Prof. César Moreno Pascual

http://es.linkedin.com/in/cesarmorenopascual/

Universidad Francisco de Vitoria
UFV Madrid

Index

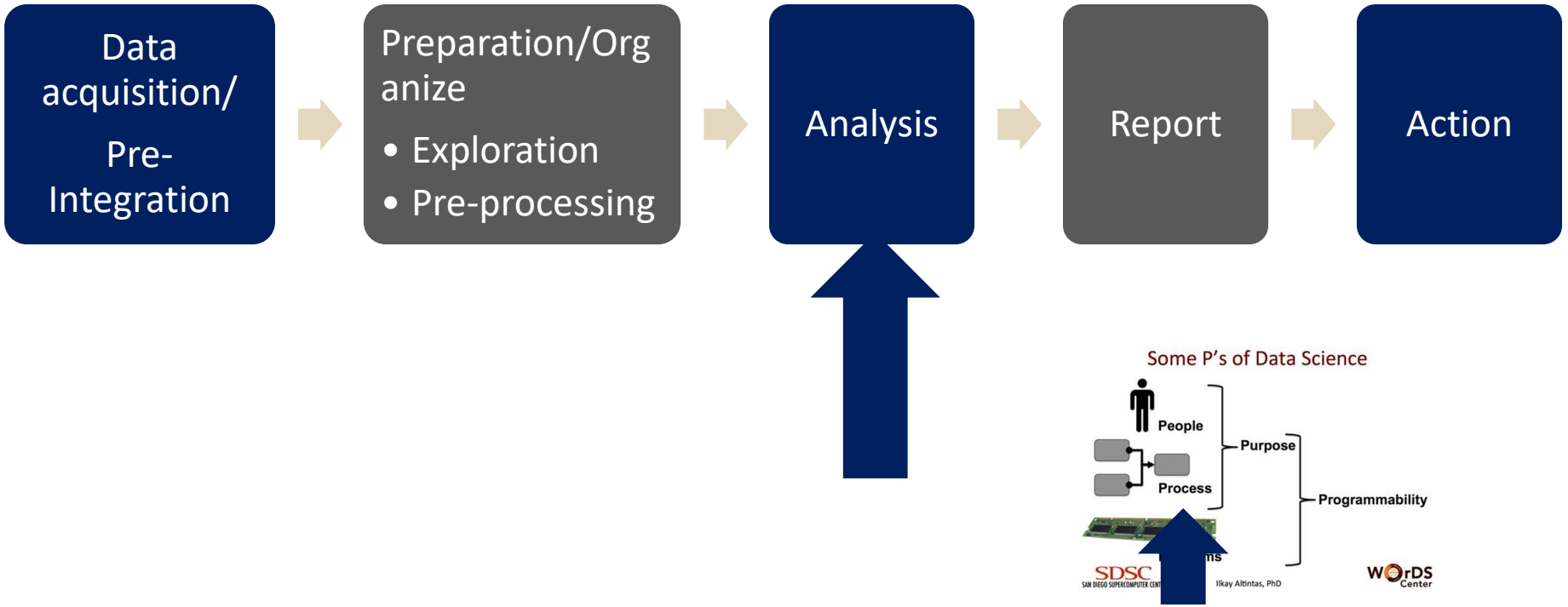
- Introduction
- Supervised
  - Prediction
  - Classification
- Unsupervised
- Structured and Unstructured data
- Deep Learning
- Other perspectives
  - Recommendation systems
  - Networks analysis

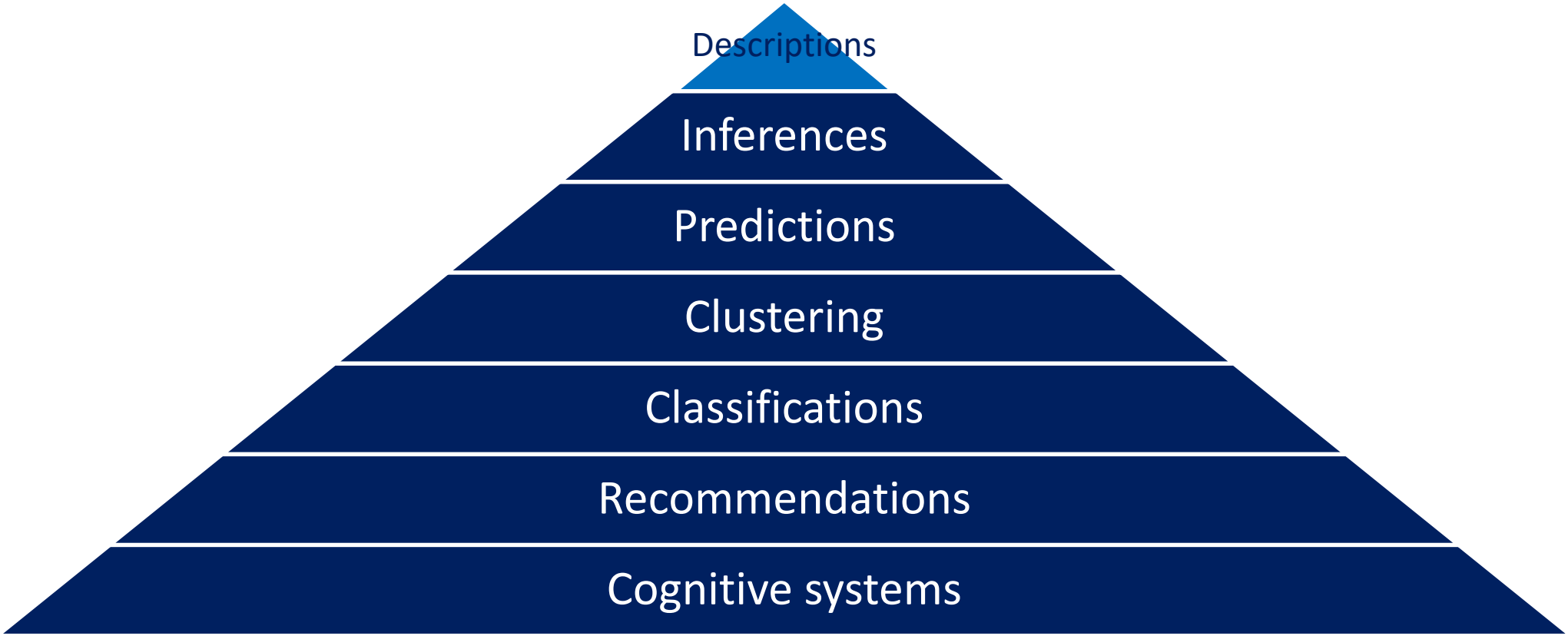


5 Ps of Big Data

Process: Remember. Cost, plan, work packages, deliverables,.....

It is a R&D project




Descriptions
Inferences
Predictions
Clustering
Classifications
Recommendations
Cognitive systems

Supervised Learning

- **Input variables or predictors or independent variables or features**
- **Output variables or response or dependent variables**
- More generally, suppose that we observe a quantitative response Y and p different predictors, X1, X2, . . . , Xp. We assume that there is some relationship between Y and X = (X1, X2, . . . , Xp), which can be written in the very general form

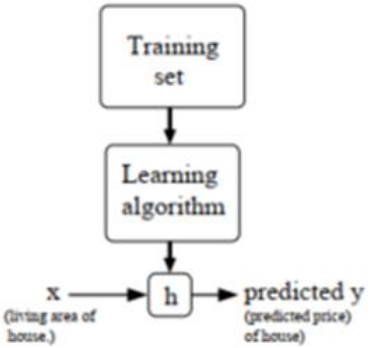
$$Y = f(X) + €$$

- **f is some fixed but unknown function of X1, . . . , Xp,**
- **€ is a random error term, which is independent of X and has mean zero**

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

- We can generally divide the techniques in

- **Supervised Learning**
  - **we have examples of inputs and outputs associated with each other:**
    - **Regression (prediction and inference)**
    - **Logistic Regression (classification)**
    - **KNN (prediction and classification)**
    - **General Additive models (prediction and classification)**
    - **Naïve Bayes and Bernoulli (Classification)**
    - **Support Vector machines (Classification)**
    - **Trees (prediction and classification)**
    - **Neural Networks (prediction and classification)**



- **Unsupervised learning**
  - we have some measures but not associated with the response or output. In this situation, we seek to know the relationship between observations
    - **Vector space classification, K-Nearest Neighbours**
    - **Principal components**

Supervised Learning; Regression (Prediction –Inference); Input and output varibles

- In a standard machine learning algorithm, we have two different types of variables:
  - **Input variable**: typically denoted as Xi with a subscript are sometimes called predictors or independent variables
  - **Output variables**: usually denoted as Y, often called response or dependent variable
- We suppose some relationship between the input and the output and write

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

  - **f** is the real function that relates X with Y, and it contents the systematic information
  - **ε** is a random error
    - that must be independent of X, nonrelated
    - must have mean zero, the distribution must be symmetrical around zero

Therefore the issue is to estimate f and evaluate its performance for two possible purposes:

- **Prediction**
- **Inference**

Supervised Learning; Regression (Prediction) Reducible and irreducible error

- **Regression (Prediction –Inference)**
  - **Prediction**: since the term **averages 0 we can predict using** $\hat{f}$ **as a black box:**

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

  - **The Prediction** $\hat{Y}$ **of Y depends on two quantities:**
    - **Reducible error**
      - It depends on the accuracy of $\hat{f}$
    - **Irreducible error**
      - Remember that Y depends also on Ɛ that cannot be predicted using X
        - The error contains unmeasured variables

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

- **Regression (Prediction)**
  - **Prediction**: since the term **averages 0 we can predict using** $\hat{f}$ **as a black box:**

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

  - **The Prediction** $\hat{Y}$ **of Y depends on two quantities:**
    - **Reducible error**
      - It depends on the accuracy of $\hat{f}$
    - **Irreducible error**
      - Remember that Y depends also on Ɛ that cannot be predicted using X
        - The error contains unmeasured variables

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

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

- **Regression (Inference)**
  - Inference: we are interested in understanding the way that Y is affected as X: X_1…….X_p change

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

  - **Now $\hat{f}$ is not a black box because we need to know its exact form to:**
    - Which predictors are associated with the response
    - What the relationship between the response and each predictor is

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning; (Prediction –Inference): parameters estimation

- **(Prediction –Inference): parameters estimation**
  - **How do we estimate $\hat{f}$:Parametric methods**
    - **Step 1: we assume a functional form or shape of**

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

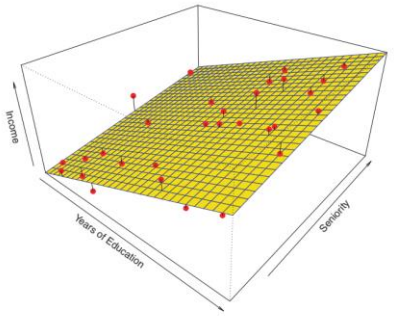


FIGURE 2.4. *A linear model fit by least squares to the* Income *data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.*

    - ***Step 2: we use the training data to fit or train the model***
      - ***We estimate the parameters***
        - Ordinary least squares: gradient descent
  - **The model we choose will usually not match the true unknown *f***
    - ***We can choose flexible models that can fit many different functional forms but:***
      - The more complex supposes to calculate more parameters
      - Can lead to overfitting

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning; Non parametric methods

- **How do we estimate $\hat{f}$: Non parametric methods**
  - Do **not make explicit assumptions** about the function form of ***f***
  - We seek to estimate f that's gets as close to the data points as possible without being too rough or wiggly
  - Parametric tests assume underlying statistical distributions in the data. Therefore, several conditions of validity must be met so that the result of a parametric test is reliable. For example, Student's t-test for two independent samples is reliable only if each sample follows a normal distribution and if sample variances are homogeneous.
  - Nonparametric tests do not rely on any distribution. They can thus be applied even if parametric conditions of validity are not met
- Since we do not reduce the calculation to a small number of parameters, **we need a very large number of observations**

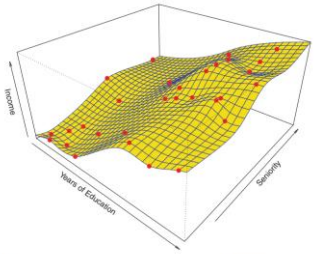


FIGURE 2.6. *A rough thin-plate spline fit to the* Income *data from Figure 2.3. This fit makes zero errors on the training data.*

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

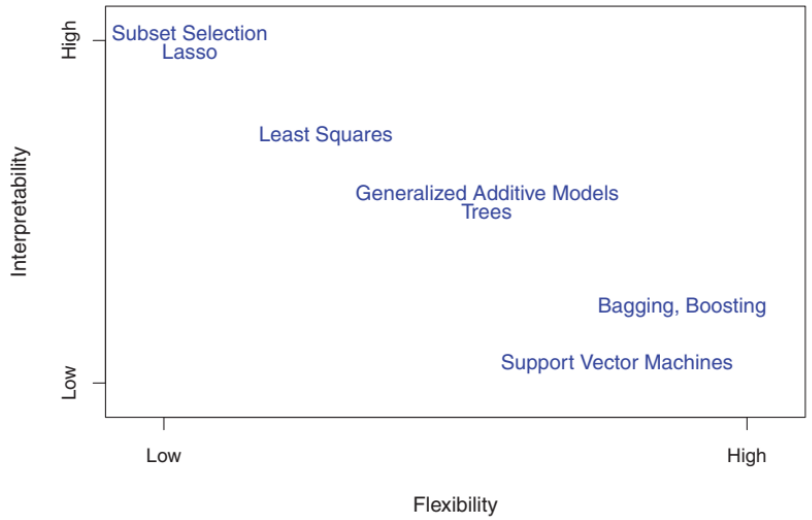
Supervised Learning; Prediction accuracy and Model interpretability

- **Prediction accuracy and Model interpretability**
  - Some **less flexible methods** as linear regression can produce a relatively small range of shapes to estimate f
  - Others, **more flexible**, as thin plate splines can generate a much **wider range of possible shapes**
  - ***Why would we ever choose to use a more restrictive method instead of a very flexible approach?***

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning; Prediction accuracy and Model interpretability

- **Prediction accuracy and Model interpretability**
  - ***Why would we ever choose to use a more restrictive method instead of a very flexible approach?***
    - ***For inference:*** more flexible are less interpretable



  - ***F*** ...ons using a less flexible
    m

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning:Prediction accuracy and Model interpretability

- **Prediction accuracy and Model interpretability**
  - ***Trade-off Variance-bias***
    - **Variance**: refers to the **amount by which $\hat{f}$ would change** if we estimated it using **different data set (the shape of $\hat{f}$ doesn't change)**
      - Since the training data are u$\hat{f}$d to fit the $\hat{f}$, **different data sets will result in a different $\hat{f}$**
      - Ideally the estimate do not vary too much $\hat{f}$ it, if a method has **high variance** then smal$\hat{f}$hanges can result in **large changes**
    - ***Bias:*** refers to the error that is introduced by **approximating a real-life problem by a much simple model (the change because we change $\hat{f}$)**
  - **In more flexible methods, the variance will increase and tl$\hat{f}$ bias will decrease**

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning:Prediction accuracy and Model interpretability

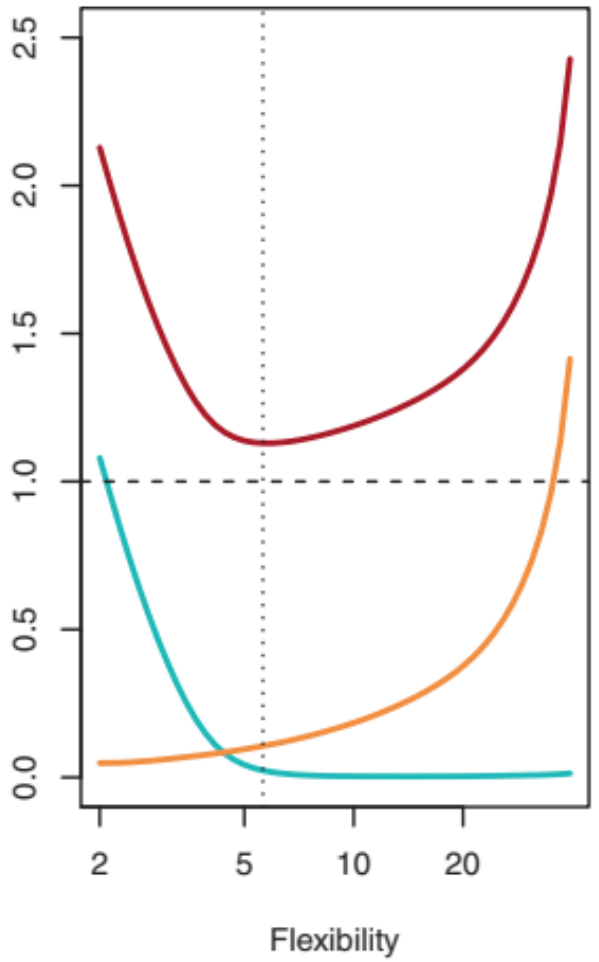
- *Trade-off Variance-bias*

- **More flexible methods the variance will increase and the bias will decrease**

  - **Orange: variance (because the change in training data)**

  - **Blue: bias (because the type of model)**

  - **Red : Least Square error (measure of the accuracy of the method)**



An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

- **Regression: Linear models_Prediction and Inference**
  - **Simple Linear regression**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \qquad h_\Theta = \Theta_0 + \Theta_1 X,$$

  - **Estimating coefficients: Least squares, gradient descent**

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$j = 0 : \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)$$

$$j = 1 : \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x^{(i)}$$

  - **Accuracy of the method: alternatives**
    - **Hypothesis test**
      - **H_0:** there is no relationship between the predictor and the response
      - **H_1:** there is relationship between the predictor and the response
        - We compute the **t-statistic**. The t-distribution is the probability of observing the value t or larger assuming the parameter of the model zero. This probability is p-value, if p-value is very small then the model is ok

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

    - **R^2:** provides the proportion of variance explained taking a value between 0 and 1

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning; Extension of Linear models_Prediction and Inference

- **Regression: Extension of Linear models_Prediction and Inference**
  - **Multiple Linear regression: multiple predictors**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

  - **Qualitative predictors**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

  - **Extensions of the linear model**
    - **Interactions**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

    - **Non-linear relationships**

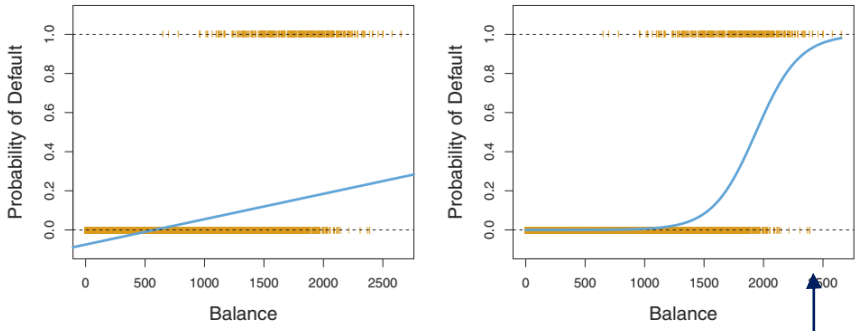
$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning; Classification: Linear models

- **Logistic Regression: to classify a response 0 or 1 linear regression is not adequate, so we model the probability of being in one group or the other instead**

$$p(X) = \beta_0 + \beta_1 X.$$



- **But this approach is not sensible eno**
  no. so to have a continuous response we finally calculate

$$\hat{y} = \sigma(w^T x + b), \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

- **Estimating coefficients: no least squares but other loss function and then gradient descent to apply it**

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Andrew Neg_deeplearning.ie
Own elaboration

Supervised Learning; Classification: Naïve Bayes classifier

- **Classification: Naïve Bayes classifier**

- **Let apply this method to text classification. The probability of a document being in class c is computed as**

$$\hat{c} = \max_{c \in C} \Pr(c) \prod_{i=1}^{n} \Pr(f_i \mid c)$$

  - **$P(c_i)$ is the probability a training set document is in class c_i. To calculate $P(c_i)$:**

$$\begin{aligned} \Pr(c_i) &= \frac{\text{number of docs of class } c}{\text{total number of docs in training dataset}} \\ &= \frac{N_c}{N_{docs}} \end{aligned}$$

  - **$P(w_i|c_i)$ is the fraction of times word w_i appears in all documents of class ci. First, we create a vocabulary V of unique words in our training set**

$$\begin{aligned} \Pr(w_i|c) &= \frac{\text{number of times } w_i \text{ appears in docs of class c}}{\text{total number of words in class c in training dataset}} \\ &= \frac{count(w_i, D_c)}{\sum_{w' \in V} count(w', D_c)} \\ &= \frac{count(w_i, D_c)}{\sum_{d \in D_c} len(d)} \quad \text{more intuitive sum} \end{aligned}$$

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V}(T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B'}$$

Introduction to information retrieval
Own elaboration

Supervised Learning:Other classification methods_Classification

- **Other classification methods_Classification**
  - **Multi-Logistic Regression: binary response with multiple predictors**
  - **Logistic regression with more than 2-classes**
  - **Naïve Bayes Classifier**
  - **Linear discriminant Analysis**
  - **K-nearest Neighbours**
  - **Trees**
  - **Neural Networks**

- **Non linear methods: prediction and classification**

  - **Polynomial regression**

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i,$$

  - **Regression splines: instead of fitting a high-degree polynomial we fit a low-degree polynomial and we smooth the connexions**

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

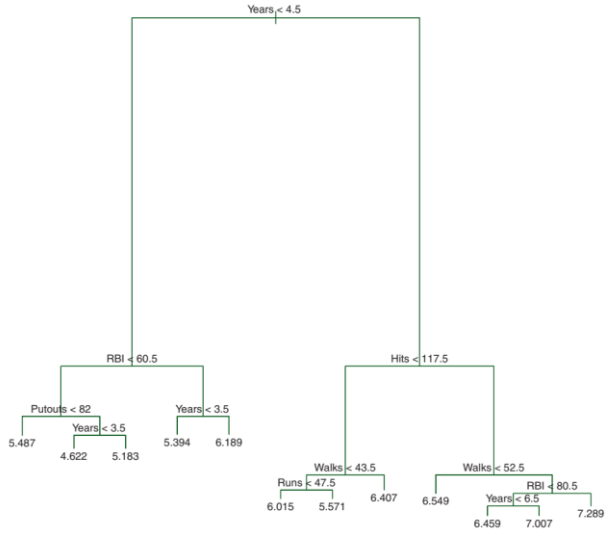
  - **Generalized additive models: general framework for extending a linear model. Now the predictors are functions**

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{\cdot} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i. \end{aligned}$$

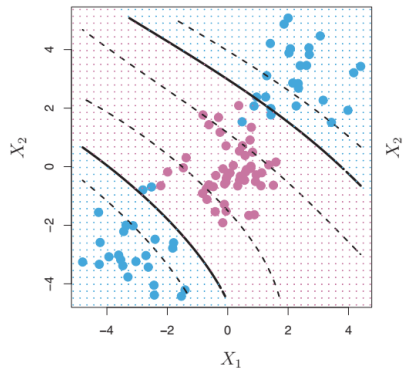
An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Supervised Learning; Tree based methods: prediction and classification

- **Tree based methods: prediction and classification**
  - **Building a tree (approx. Method)**
    - We divide the predictor into some nob-overlapping regions
    - For every observation that falls in the same region we make the same prediction
    - We apply a cost function (cos complexity pruning function)
    - Repeat until the division is optimal



- **Support Vector Machines: Classification**
  - Is a generalization of other called maximal margin classifier
    - Classifies using hyperplanes



An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

- **Principal components**
  - The principal components approach involves **constructing principal components** and then using these **components as predictors in a linear regression model** that is fit using least squares
  - Often a small number of components suffice to explain most of the variability
  - We assume **the directions in which the predictors X show most variation** are the directions that **are associated with Y**

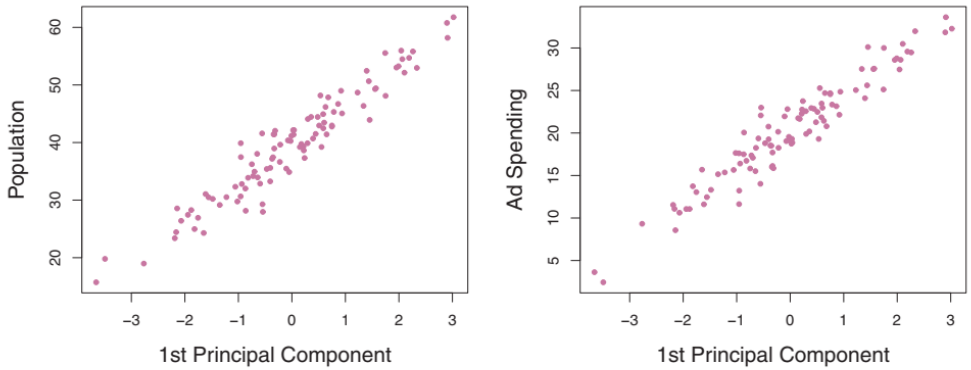


FIGURE 6.16. *Plots of the first principal component scores z_{i1} versus* pop *and* ad. *The relationships are strong.*

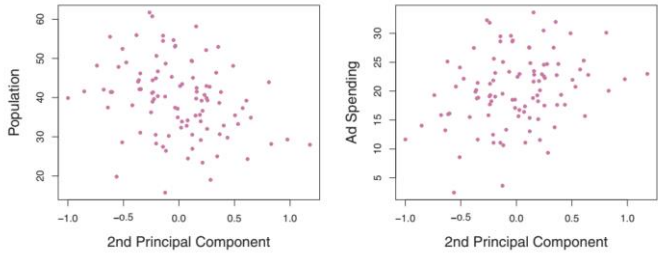


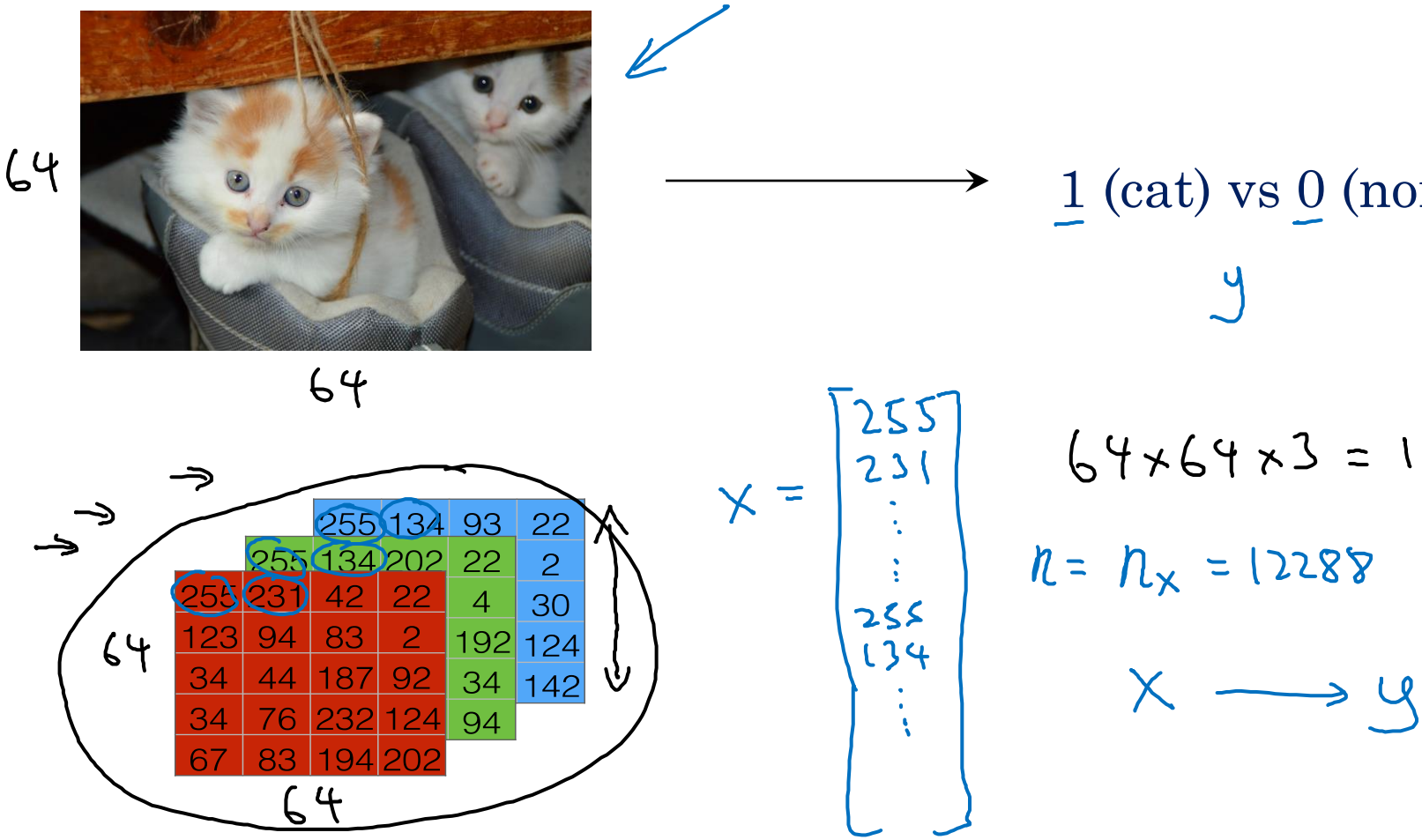
FIGURE 6.17. *Plots of the second principal component scores z_{i2} versus* pop *and* ad. *The relationships are weak.*

In the advertising data, the first principal component explains most of the variance in both pop and ad, so a principal component regression that uses this single variable to predict some response of interest, such as sales, will likely perform quite well.

An introduction to statistical learning _ James G.,Witten,D.,Hastie,T,Tibshirani R._Springer New Yor Hiedelberg Dordrecht London
Own elaboration

Structured and unstructured Data; Unstructured Data: Vector space model

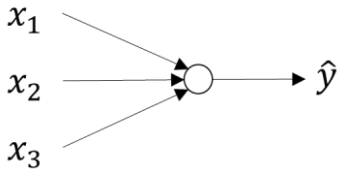
- **Unstructured Data: Vector space model**

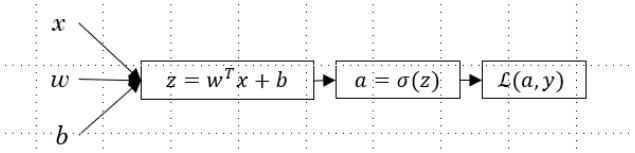


Andrew Ng_deeplearning.ie
Own elaboration

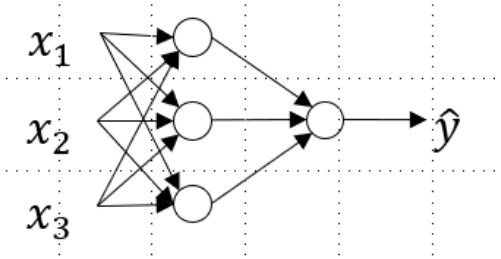
Deep Learning (Neural networks)

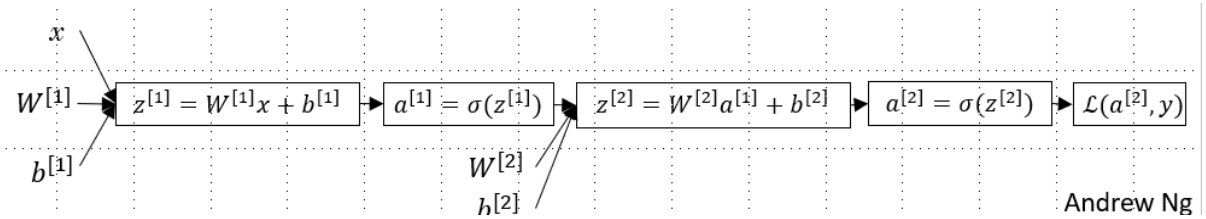
- **Deep Learning: regression and classification (prediction)**
  - ***For example a Logit Classification is a one neuron network: this is the computation network***





  - ***A Neuron Network: a full connected network with several layers.***
    - In the middle we find the **hidden layers** that there are **calculated automatically**.
    - The next layers are calculated using as an input the output of the previous layer





Andrew Ng_deeplearning.ie
Own elaboration

Neural Networks: Supervised Learning usage

Input(x)	Output (y)	Application
Home features	Price	Real Estate
Ad, user info	Click on ad? (0/1)	Online Advertising
Image	Object (1,…,1000)	Photo tagging
Audio	Text transcript	Speech recognition
English	Chinese	Machine translation
Image, Radar info	Position of other cars	Autonomous driving

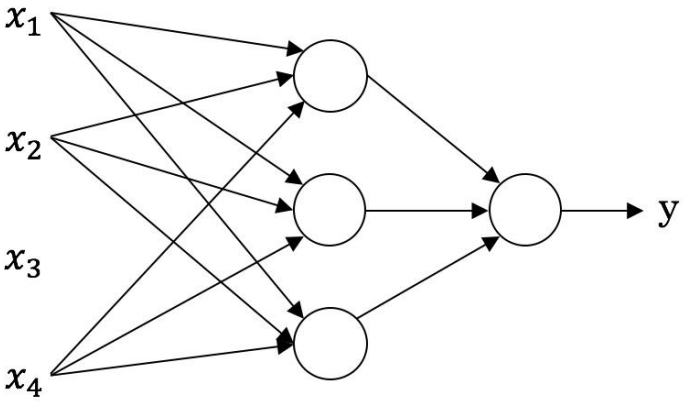
Standard NN

CNN

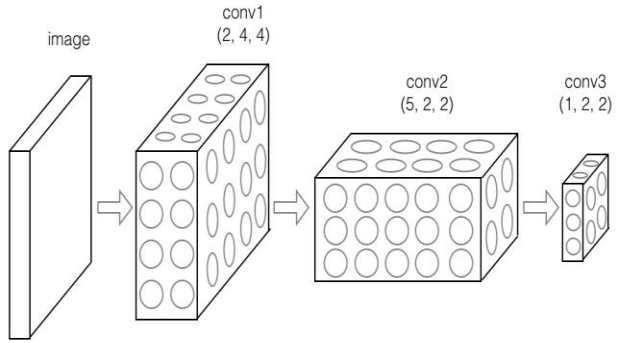
RNN

Custom/ Hybrid

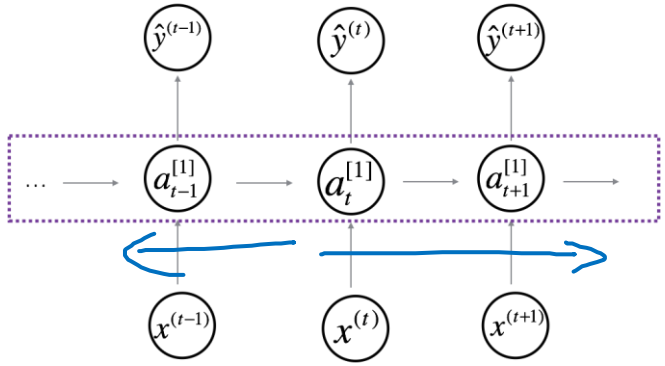
Neural Network examples: supervised and unsupervised applications



Standard NN



Convolutional NN



Recurrent NN

Other perspectives

- **There are many other techniques in this subject but also some other categories for specific purposes as Recommendation systems:**
  - there is a mixed category as it uses some of the previous techniques
  - Types:
    - Non-personalized summary statistics
    - Content based
    - Collaborative filtering
      - User-User
      - Item-Item
      - Dimensionality reduction

- **Also one other completely different approach are Networks**
  - Networks that represent real underlaying relationships (social, economic or any other type)
  - There are different to computational Networks
  - *Google Search uses as part of its engine this approach : Pagerank*

Andrew Ng_deeplearning.ie
Own elaboration

BOOKs READINGS APPLIED TO THIS PRESENTATION

Book2: Introduction to statistical learning (James, Witten)

- Chapter 2

Book3: Introduction to information retrieval (Manning , Raghavan,Schüzte)

- Chapter 13