Lecture 3c. Basis Expansion.

COMP90051 Statistical Machine Learning

Semester 2, 2020
Lecturer: Ben Rubinstein

THE UNIVERSITY OF MELBOURNE

This lecture

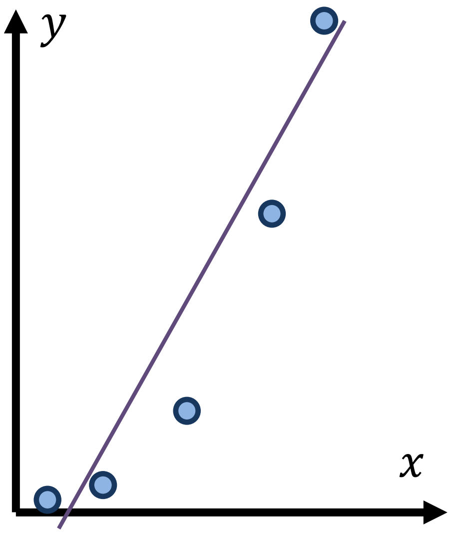
- Linear regression
  - Simple model (convenient maths at expense of flexibility)
  - Often needs less data, "interpretable", lifts to non-linear
  - Derivable under all Statistical Schools: Lect 2 case study
    - This week: Frequentist + Decision theory derivations
    - Later in semester: Bayesian approach
  - Convenient optimisation: Training by "analytic" (exact) solution
- **Basis expansion: Data transform for more expressive models**

Basis Expansion

Extending the utility of models via data transformation

Basis expansion for linear regression

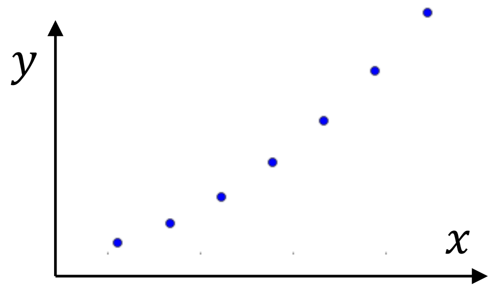
- Real data is likely to be non-linear
- What if we still wanted to use a linear regression?
    - Simple, easy to understand, computationally efficient, etc.
- How to marry non-linear data to a linear method?



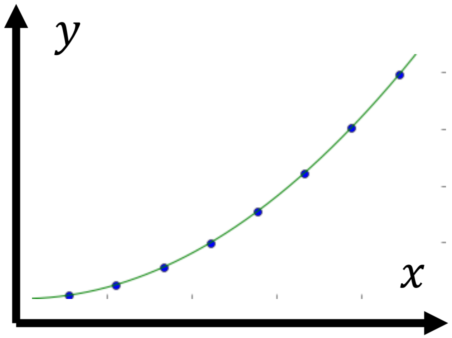
If you can't beat'em, join'em

Transform the data

- The trick is to transform the data: Map data into another features space, s.t. data is linear in that space
- Denote this transformation $\varphi: \mathbb{R}^m \rightarrow \mathbb{R}^k$. If $\boldsymbol{x}$ is the original set of features, $\varphi(\boldsymbol{x})$ denotes new feature set
- Example: suppose there is just one feature x, and the data is scattered around a parabola rather than a straight line



Example: Polynomial regression



- No worries, mate: define

$$\varphi_1(x) = x$$
$$\varphi_2(x) = x^2$$

- Next, apply linear regression to φ_1, φ_2

$$y = w_0 + w_1\varphi_1(x) + w_2\varphi_2(x) = w_0 + w_1 x + w_2 x^2$$

and here you have **quadratic regression**

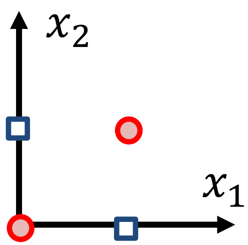
- More generally, obtain **polynomial regression** if the new set of attributes are powers of x
- Similar idea basis of autoregression for time series

Example: linear classification

- Example binary classification problem: Dataset not linearly separable
- Define transformation as

$$\varphi_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{z}_i\|, \text{ where } \boldsymbol{z}_i \text{ some pre-defined constants}$$

- Choose $\boldsymbol{z}_1 = [0,0]'$, $\boldsymbol{z}_2 = [0,1]'$, $\boldsymbol{z}_3 = [1,0]'$, $\boldsymbol{z}_4 = [1,1]'$



x_1	x_2	y
0	0	Class A
0	1	Class B
1	0	Class B
1	1	Class A

there exist weights that make new data separable, e.g.:

w_1	w_2	w_3	w_4
1	0	0	1

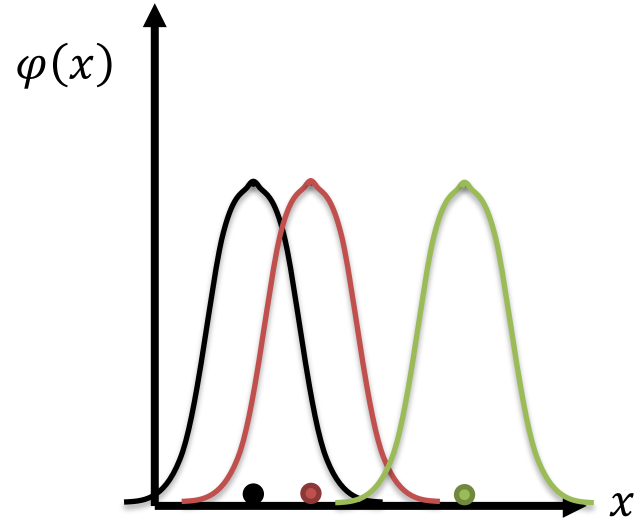
φ_1	φ_2	φ_3	φ_4
0	1	1	$\sqrt{2}$
1	0	$\sqrt{2}$	1
1	$\sqrt{2}$	0	1
$\sqrt{2}$	1	1	0

The transformed data is linearly separable!

$\boldsymbol{\varphi}'\boldsymbol{w}$	y
$\sqrt{2}$	Class A
2	Class B
2	Class B
$\sqrt{2}$	Class A

Radial basis functions

- Previous example: motivated by approximation theory where sums of RBFs approx. functions
- A radial basis function is a function of the form $\varphi(\boldsymbol{x}) = \psi(\|\boldsymbol{x} - \boldsymbol{z}\|)$, where $\boldsymbol{z}$ is a constant



- Examples:
- $\varphi(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{z}\|$
- $\varphi(\boldsymbol{x}) = \exp\left(-\frac{1}{\sigma}\|\boldsymbol{x} - \boldsymbol{z}\|^2\right)$

Challenges of basis expansion

- Basis expansion can significantly increase the utility of methods, especially, linear methods
- In the above examples, one limitation is that the transformation needs to be defined beforehand
  - Need to choose the size of the new feature set
  - If using RBFs, need to choose $\boldsymbol{z}_i$
- Regarding $\boldsymbol{z}_i$, one can choose uniformly spaced points, or cluster training data and use cluster centroids
- Another popular idea is to use training data $\boldsymbol{z}_i \equiv \boldsymbol{x}_i$
  - E.g., $\varphi_i(\boldsymbol{x}) = \psi(\|\boldsymbol{x} - \boldsymbol{x}_i\|)$
  - However, for large datasets, this results in a large number of features → computational hurdle

Further directions

- There are several avenues for taking the idea of basis expansion to the next level
  - Will be covered later in this subject
- One idea is to *learn* the transformation φ from data
  - E.g., Artificial Neural Networks
- Another powerful extension is the use of the kernel trick
  - "Kernelised" methods, e.g., kernelised perceptron
- Finally, in sparse kernel machines, training depends only on a few data points
  - E.g., SVM

Summary

- Basis expansion
  - Extending model expressiveness via data transformation
  - Examples for linear and logistic regression
  - Theoretical notes

Next time:

First/second-order iteration optimisation;

Logistic regression - linear probabilistic model for classification.