Lecture 3b. Linear Regression - Frequentist.

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

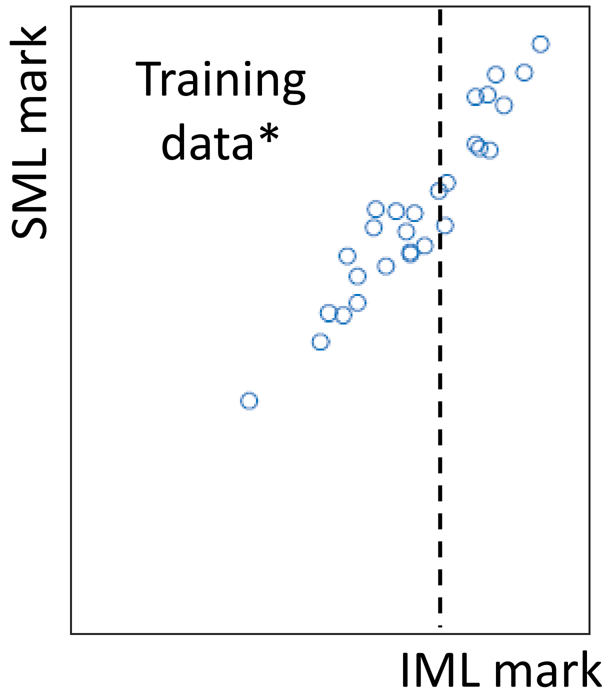
- Basis expansion: Data transform for more expressive models

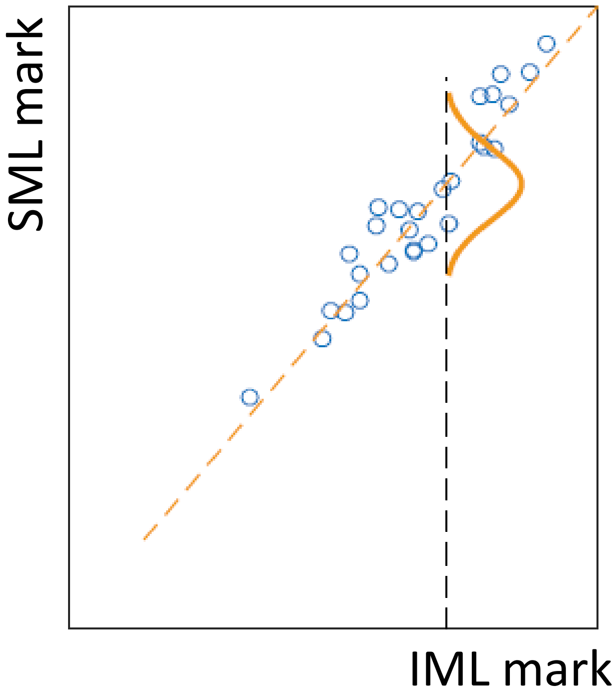
Linear Regression via Frequentist Probabilistic Model

Max-Likelihood Estimation

Data is noisy!

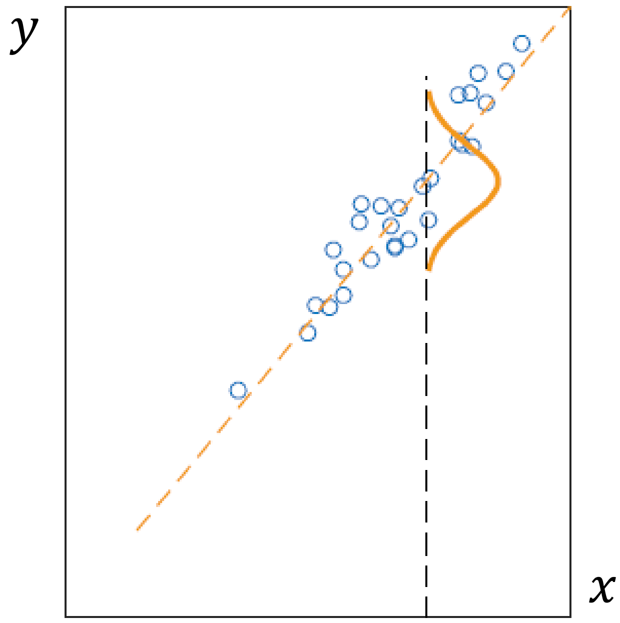
Example: predict mark for Statistical Machine Learning (SML) from mark for Intro ML (IML aka KT)





* synthetic data :)

Regression as a probabilistic model



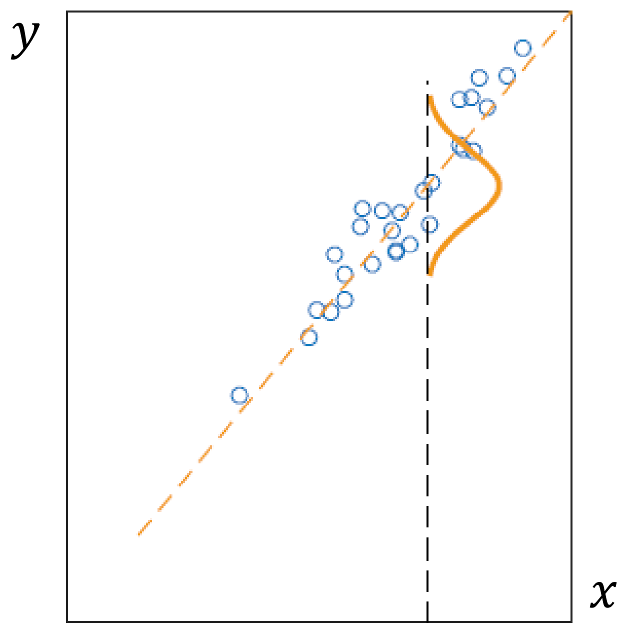
- Assume a probabilistic model: $Y = \boldsymbol{X}'\boldsymbol{w} + \varepsilon$
  - Here $\boldsymbol{X}, Y$ and ε are r.v.'s
  - Variable ε encodes noise
- Next, assume Gaussian noise (indep. of $\boldsymbol{X}$): $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

- Recall that $\mathcal{N}(x; \mu, \sigma^2) \equiv \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- Therefore

$$p_{\boldsymbol{w},\sigma^2}(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\boxed{(y - \boldsymbol{x}'\boldsymbol{w})^2}}{2\sigma^2}\right)$$

this is a squared error!

Parametric probabilistic model



- Using simplified notation, discriminative model is:

$$p(y|\boldsymbol{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \boldsymbol{x}'\boldsymbol{w})^2}{2\sigma^2}\right)$$

- Unknown parameters: $\boldsymbol{w}, \cancel{\sigma^2}$

- Given observed data $\{(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_n, Y_n)\}$, we want to find parameter values that "best" explain the data

- Maximum-likelihood estimation: choose parameter values that maximise the probability of observed data

Maximum likelihood estimation

- Assuming independence of data points, the probability of data is

$$p(y_1, \dots, y_n | \boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \prod_{i=1}^{n} p(y_i | \boldsymbol{x}_i)$$

- For $p(y_i|\boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{x}_i{}'\boldsymbol{w})^2}{2\sigma^2}\right)$

- "Log trick": Instead of maximising this quantity, we can maximise its logarithm (Why? Explained soon)

$$\sum_{i=1}^{n} \log p(y_i|\boldsymbol{x}_i) = -\frac{1}{2\sigma^2} \boxed{\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i{}'\boldsymbol{w})^2} + C$$

the sum of squared errors!

here C doesn't depend on $\boldsymbol{w}$ (it's a constant)

- Under this model, maximising log-likelihood as a function of $\boldsymbol{w}$ is equivalent to minimising the sum of squared errors

Method of least squares

Analytic solution:
- Write derivative
- Set to zero
- Solve for model

- Training data: $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$. Note bold face in $\boldsymbol{x}_i$
- For convenience, place instances in rows (so attributes go in columns), representing training data as an $n \times (m+1)$ matrix $\boldsymbol{X}$, and n vector $\boldsymbol{y}$
- Probabilistic model/decision rule assumes $\boldsymbol{y} \approx \boldsymbol{X}\boldsymbol{w}$
- To find $\boldsymbol{w}$, minimise the sum of squared errors

$$L = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} X_{ij} w_j \right)^2$$

$$L = \|y - Xw\|_2^2$$

$$\frac{\partial L}{\partial w} = \cancel{2} X'(y - X\hat{w}) = 0$$

- Setting derivative to zero and solving for $\boldsymbol{w}$ yields

$$\hat{\boldsymbol{w}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$X'y = X'X\hat{w}$$

  - This system of equations called the normal equations
  - System is well defined only if the inverse exists

In this slide: UPPERCASE symbol in bold face means matrix; $\boldsymbol{X}'$ denotes transpose; $\boldsymbol{X}^{-1}$ denotes matrix inverse

Wherefore art thou: Bayesian derivation?

- Later in the semester: return of linear regression
- Fully Bayesian, with a posterior:
  - Bayesian linear regression
- Bayesian (MAP) point estimate of weight vector:
  - Adds a penalty term to sum of squared losses
  - Equivalent to L_2 “regularisation” to be covered next week
  - Called: ridge regression

Summary

- Linear regression
  - Simple, effective, “interpretable”, basis for many approaches
  - Probabilistic frequentist derivation
  - Solution by normal equations

  Later in semester: Bayesian approaches

Next time: Basis expansion for non-linear regression