COMP90051
Statistical Machine Learning

Workshop Week 4

Xudong Han

https://github.com/HanXudong/COMP90051_Workshops

Your tutor

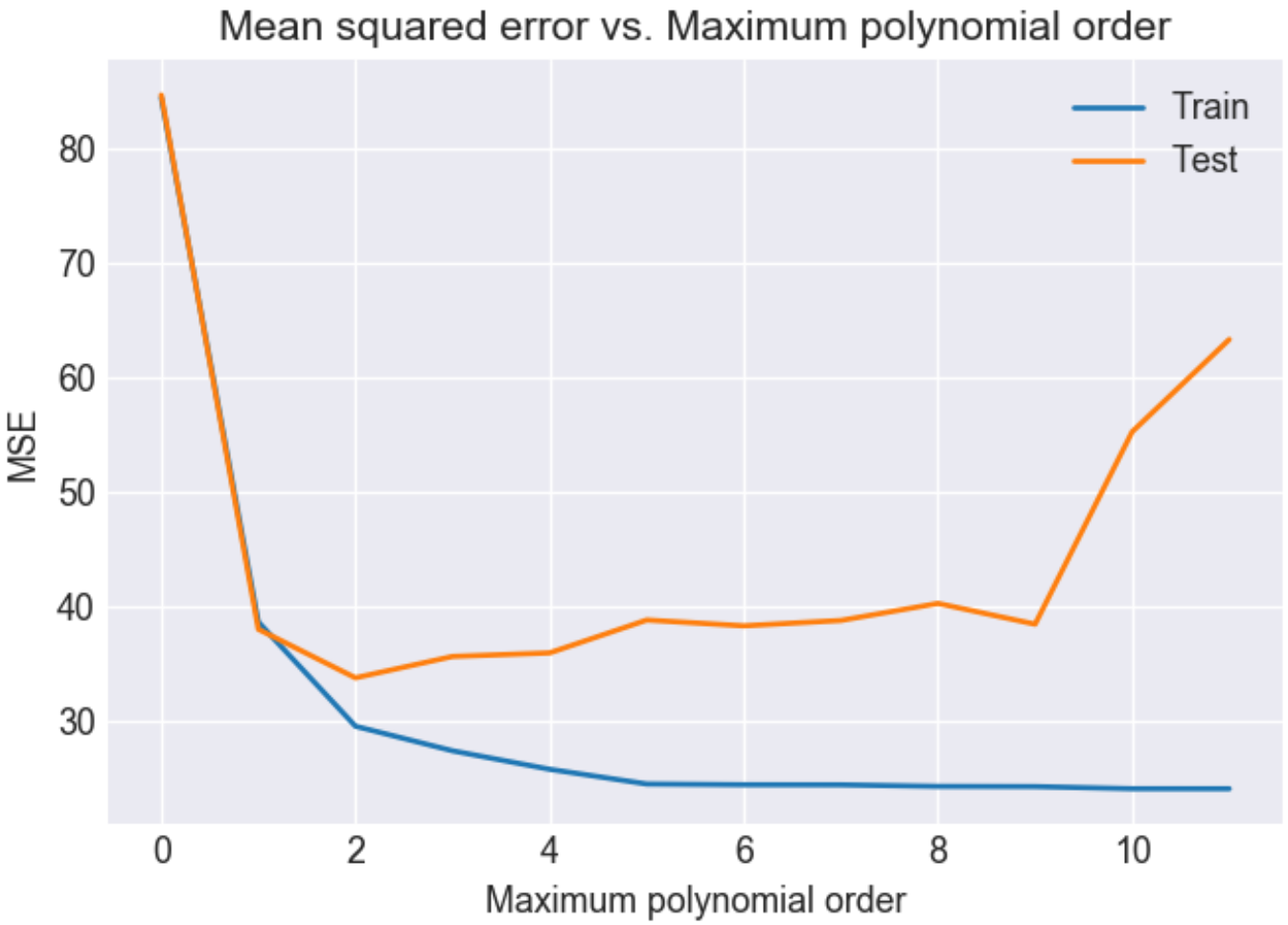
- Xudong Han
- xudong.han@unimelb.edu.au

- Slides https://github.com/HanXudong/COMP90051_Workshops

Review

- Linear regression: $y = \vec{x}^T \vec{\beta} + \varepsilon$
  - $y \sim N(\vec{x}^T \vec{\beta}, \sigma^2)$
  - $\vec{x}^T = \begin{pmatrix} 1 & x_1 & x_2 & \ldots. & x_p \end{pmatrix}$
  - $\vec{\beta}^T = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \ldots. \quad \beta_p)$
  - $\varepsilon \sim N(0, \sigma^2)$
- Maximum Likelihood Estimator == Least Square Estimator: $\hat{\beta} = (X^T X)^{-1} X^T y$

Review


Mean squared error vs. Maximum polynomial order
Train
Test
MSE
80
70
60
50
40
30
0
2
4
6
8
10
Maximum polynomial order

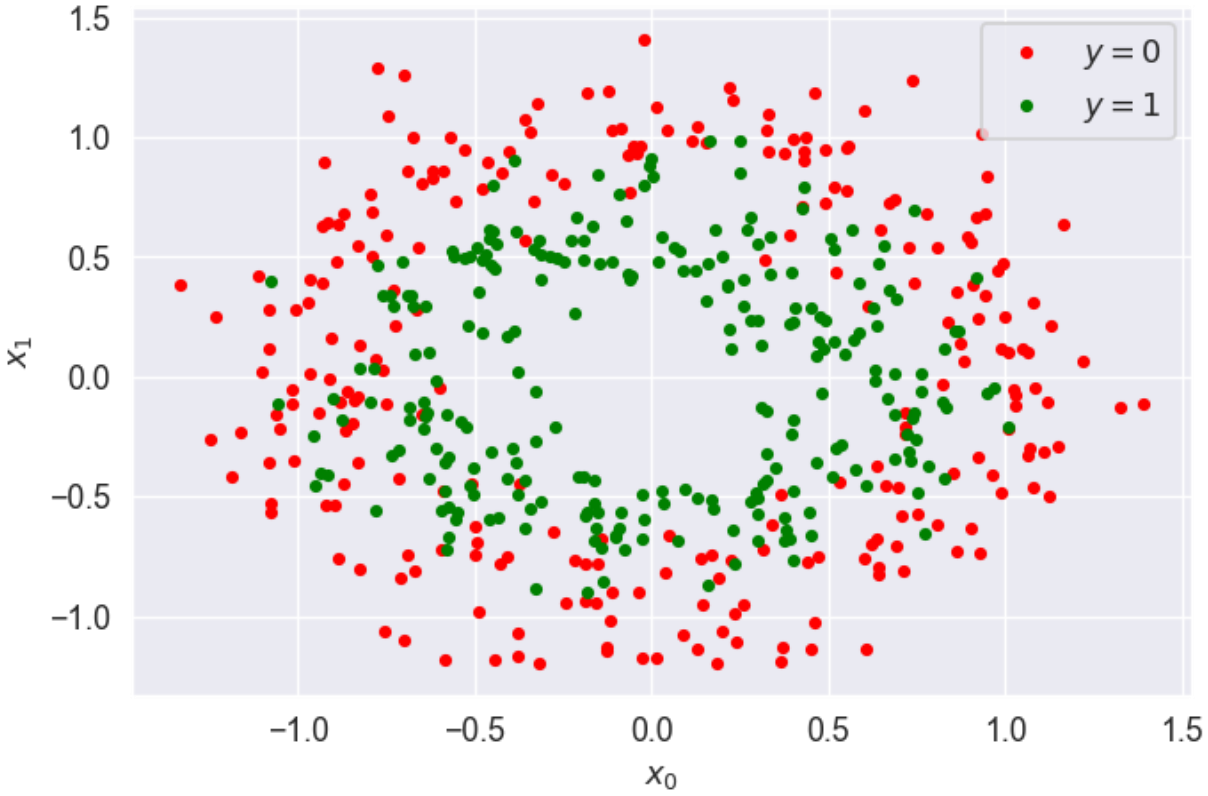
Learning Outcomes

At the end of this workshop you should:

- Be able to implement regularised logistic regression using a numerical optimisation solver
- Be able to apply basis expansion to turn logistic regression into a non-linear classifier
- Understand the purpose/effect of L2 regularisation

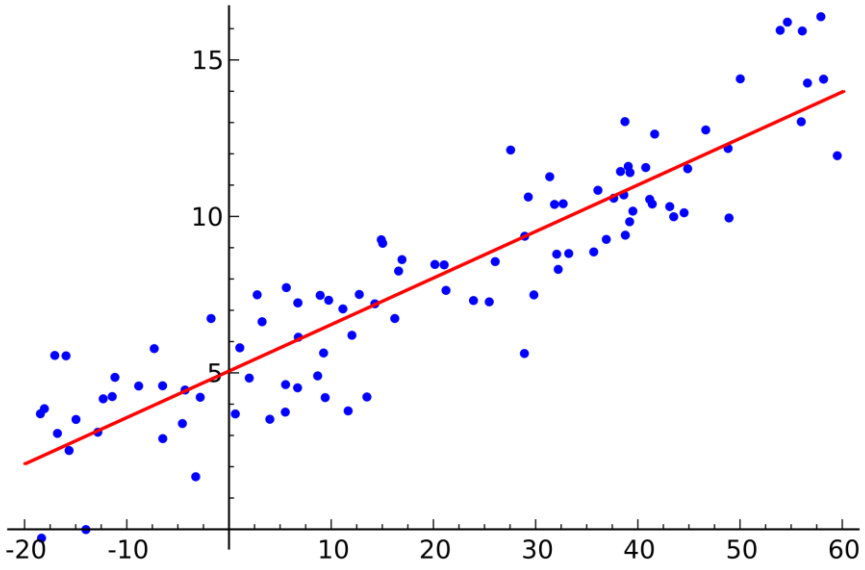
Binary classification data

- $y_n \sim Bernoulli(\theta)$
- $f(y_n|\theta) = \theta^{y_n}(1-\theta)^{(1-y_n)}$



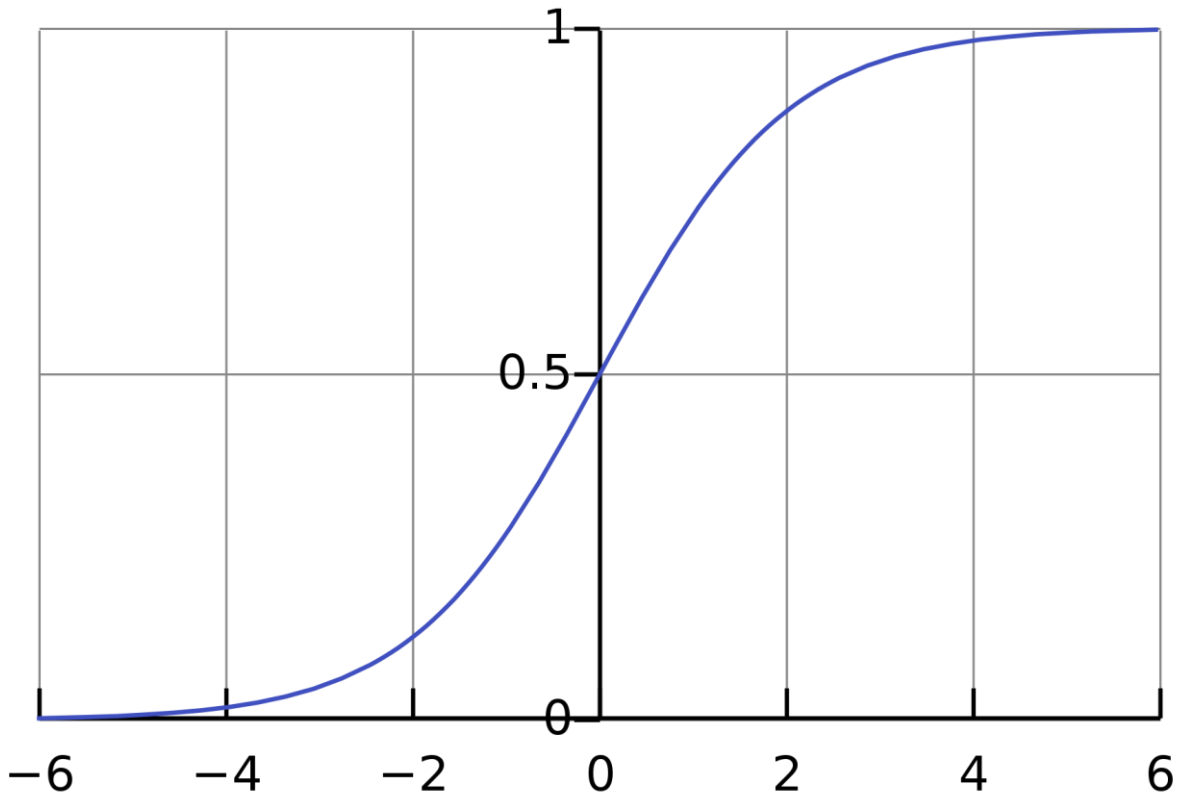
Logistic Regression

- $y|\vec{x} \sim Bernoulli[\theta]$
- The range of θ : $0 \leq \theta \leq 1$
- Compared with Linear regression



Logistic function

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + 1}$$



Logistic Regression = Linear Regression + Logistic function

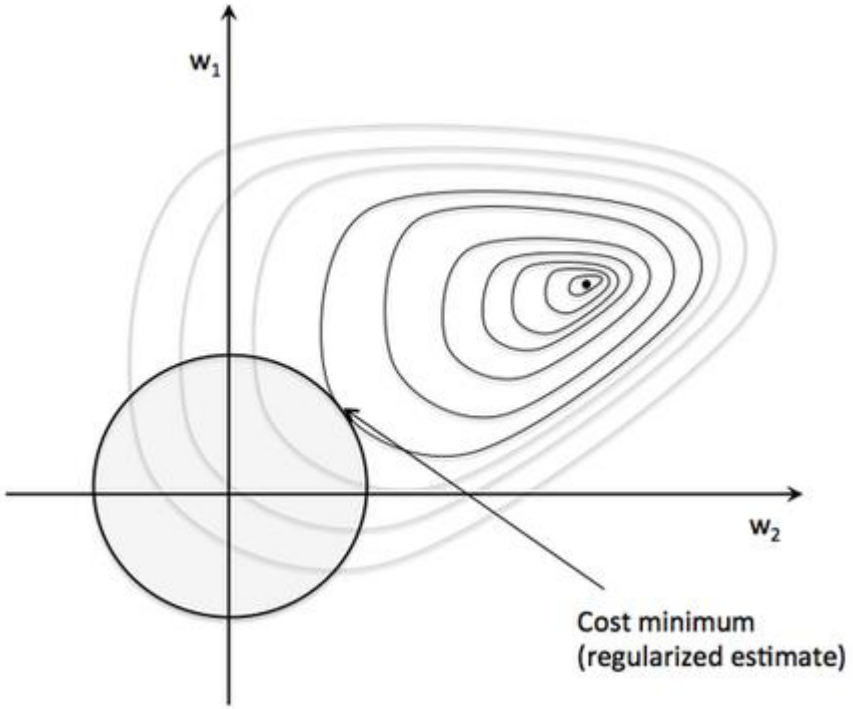
- $\theta = p(y = 1|\vec{x}) = \sigma(\vec{w}^T \vec{X} + b)$
- $\vec{w}^T = (\beta_1 \quad \beta_2 \quad \dots \quad \beta_p)$
- $\vec{X}^T = (x_1 \quad x_2 \quad \dots \quad x_p)$
- $b = \beta_0$

- Estimating parameters $[\vec{w}^T \quad b]$
- Maximize Log-likelihood == Minimize Cross-Entropy (See Jupyter notebook)

Exercise 1 (Discussion?):

- The L2 regularization term $\mathcal{L}_{reg}(w) = \frac{1}{2}\lambda\vec{w}^T\,\vec{w}$ is commonly said to reduce overfitting. Give a brief justification why?
  - What is overfitting?
  - What is the meaning of regularization term?

Regularization



$$w_1^2 + w_2^2 \leq C^2$$

https://github.com/rasbt/python-machine-learning-book/tree/master/faq

Exercise 2 (Discussion?):

- Why do we only include the weights w in the L2 regularization term? i.e. the bias terms are excluded from regularization.

- A small bias term may force the classifier to stay within the region close to the origin where the sigmoid is approximately linear. This makes it difficult to classify data where the binary classes are not linearly separable.

Exercise 3:

- $\sigma(x) = \frac{\exp(x)}{\exp(x)+1}$

- $\log \frac{p(y=1|\vec{x})}{p(y=0|\vec{x})} = ?$