Logistic Regression 1
The basics

Michael Claudius, Associate Professor, Roskilde

31.03.2020 Revised 18.10.2020

What is logistic regression?

- **A *predicative* algorithm for classification**
- **Based on probability (p), a number**
  - **in percent: 0% $\leq$ p $\leq$ 100%;**
  - **in decimal: 0 $\leq$ p $\leq$ 1**
- **Binary classification OR**
- **Multiple classes (multinomial)**

- **Give you a minute!**
- **Toss a coin. What is the probability of heads and tails (plat eller krone)?**
- **Throw a dice. What is the probability for a 6?**
- **Throw two dice a red and a green.**

- **So its predicting something; lets look at that !**

Evaluation of logistic regression?

- **Advantages**
  - **Also good for small data sets!**
  - **White box; knows in details how it works**
  - **Easy**

- **Disadvantages**
  - **Not good for big data, too slow**
  - **Wrong estimates for messy data, outliers**
  - **No missing data**
  - **Variables (features) must be independent**

Prediction

- **Prediction, y, of an instance X (X can be one feature (X_1) or many features (vector, $X_1, X_2, \ldots X_n$))**
  - **$p \geq 0.5$ => y = 1 (X is an instance of a positive class)**
  - **$p < 0.5$ => y = 0 (X is an instance of a negative class)**
- **Notice: logistic regression is predicting just 0 or 1; not a range of values (BAM)**

- **Let us watch an easy video introduction** Logistic Regression Introduction (8 minutes)
- **Before the hard stuff**

Estimation elements

- **It is all math ☺; that's it looks complicated so just keep it simple!**

Equation 4-13. Logistic Regression model estimated probability (vectorized form)

$$\hat{p} = h_{\theta}(\mathbf{x}) = \sigma(\mathbf{x}^{\intercal}\boldsymbol{\theta})$$

- **p: estimated probability**
- **h: hypothesis function based on θ: h_{θ}**
- **X: feature vector or just feature values** $X_1, X_2, \ldots X_n$
- **θ: parameter vector weights on features ($\theta_0, \theta_1, \theta_2, \ldots \theta_n$)**
- **X^T: transposed vector (columns changed to rows)**
- **$X^T\theta$: matrix multiplication (like linear regression $\theta_0 + X_1\theta_1 + X_2\theta_2 \ldots + X_n\theta_n$**
- **σ: the famous sigmoid function !**

- **A link to Wikipedia**

Sigmoid function

Equation 4-14. Logistic function

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

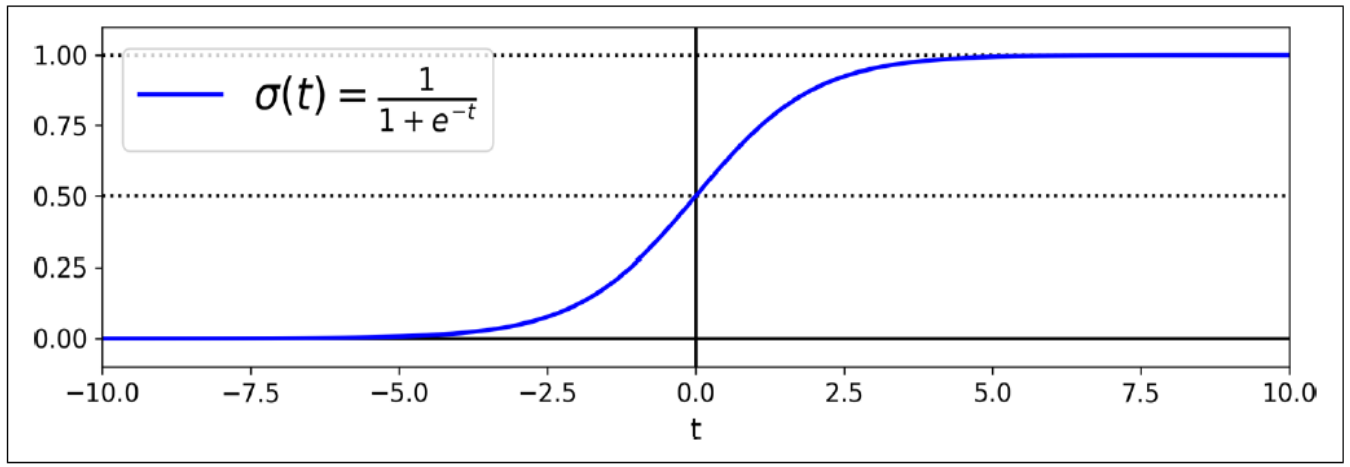

Figure 4-21. Logistic function

- **σ(t): values 0 – 1 !**

Training

- **Idea: to train the model (i.e. changing parameters $\theta_0, \theta_1, \theta_2, \ldots \theta_n$)**
- **Goal: p is high for instance of positive class and low for instances of negative class**
- **So need a cost function $c(\theta_0, \theta_1, \theta_2, \ldots \theta_n)$ fulfilling:**

  - **Cost is high for wrong estimation (false)**
    a. **Guess 0 for a positive class**
    b. **Guess 1 for a negative class**

  - **Cost is low for correct estimation (true)**
    a. **Guess 1 for a positive class**
    b. **Guess 0 for a negative class**

  - **And yes it exists! We are lucky.**

Cost function

- **This function for a single training instance fulfills the requirements**

Equation 4-16. Cost function of a single training instance

$$c(\boldsymbol{\theta}) = \begin{cases} -\log(\hat{p}) & \text{if } y = 1 \\ -\log(1 - \hat{p}) & \text{if } y = 0 \end{cases}$$

- **c: cost function**
- **θ: parameter vector weights on features (θ_0, θ_1, θ_2, ….. θ_n)**
- **p: estimated probability**

- **But of course there are many instances, so we need an average of summation…**

Average cost function

- **But of course there are many instances, so we need an average of summation of the whole training set**

Equation 4-17. Logistic Regression cost function (log loss)

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(\hat{p}^{(i)}\right) + \left(1 - y^{(i)}\right)\log\left(1 - \hat{p}^{(i)}\right)\right]$$

- **J(θ): parameter vector weights on features (θ_0, θ_1, θ_2, ….. θ_n)**

- **How to find the best set ?**
- **No Normal Equation !**
- **BUT Again we are lucky..**

Partial derivative of average cost function

- **Why Lucky?, because J(θ) is convex and differentiable**
-

Equation 4-18. Logistic cost function partial derivatives

$$\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \left(\sigma\left(\boldsymbol{\theta}^T \mathbf{x}^{(i)} \right) - y^{(i)} \right) x_j^{(i)}$$

- **That's it has a global minimum and then**
- **We can find the parameters (θ_0, θ_1, θ_2, ….. θ_n) using Batch Gradient Algorithm ! (BAM)**

Assignments

- **It is time for discussion and solving a few assignments in groups**
  - Logistic Regression Questions