Introduction to Machine Learning
Logistic Regression

Nikolay Manchev

23 March 2017

London Machine Learning Study Group

Housekeeping

Next events

http://www.meetup.com/London-Machine-Learning-Study-Group

Follow me

https://twitter.com/nikolaymanchev

Slides and code

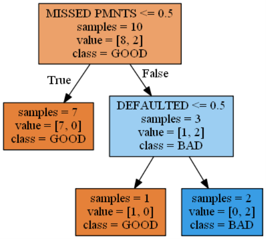
Available at https://github.com/nmanchev/MachineLearningStudyGroup

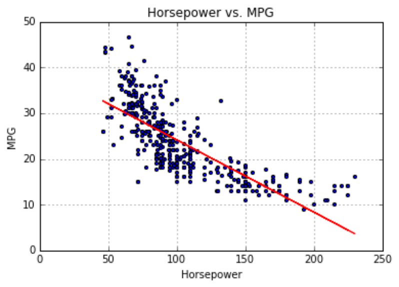
Previous recordings

Available at https://www.youtube.com/c/NikolayManchev

Flach talks about three types of Machine Learning models [Fla12]

- Geometric models
- Logical models
- Probabilistic models

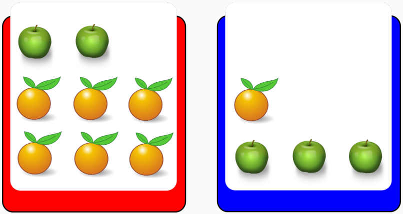




Probabilities – a refresh

Simple example [Bis06]

1. We randomly pick one of the boxes (40% probability for the red and 60% for the blue box)
2. We randomly pick a fruit



$$p(B = r) = \frac{4}{10}, p(B = b) = \frac{6}{10}$$

- By definition probabilities lie in $[0; 1]$
- If the events include all possible outcomes and are mutually exclusive their probabilities must sum to one (eg. $\frac{4}{10} + \frac{6}{10} = 1$)

Definitions

- **Marginal probability** – the probability of an event occurring is not conditioned on any other event

$$p(B = r) = \frac{4}{10}$$

- **Joint probability** – probability of the events occurring together

$$p(B = r, F = a) = ?$$

- **Conditional probability** – probability of an event occurring, given that another event occurs

$$p(B = a | B = r) = \frac{1}{4} \text{ (fraction of apples in the red box)}$$

Rules of Probability

- Sum rule – $p(X) = \sum_{Y} p(X,Y)$
- Product rule – $p(X,Y) = p(Y|X)p(X)$

Example 1 – $p(B=r, F=a) = ?$

$$p(B=r, F=a) = p(F=a|B=r) \times p(B=r) = \frac{1}{4} \times \frac{4}{10} = \frac{1}{10}$$

Example 2 – $p(F=a) = ?$

$$p(F=a) = \sum_{B \in \{r,b\}} p(F=a|B) = p(F=a|B=r) +$$

$$p(F=a|B=b) = p(F=a|B=r)p(B=r) +$$

$$p(F=a|B=b)p(B=b) = \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20}$$

Binary Logistic Regression

Definition

Binary Logistic Regression

- We have a set of feature vectors $\boldsymbol{X}$ with corresponding binary outputs

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}^\mathsf{T}$$
$$\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}^\mathsf{T}, \text{ where } y_i \in \{0, 1\}$$

- We want to model $p(y|\boldsymbol{x})$

$$p(y_i = 1|\boldsymbol{x_i}, \boldsymbol{w}) = \sum_j w_j x_{ij} = \boldsymbol{x_i}\boldsymbol{w}$$

By definition $p(y_i = 1|\boldsymbol{x_i}, \boldsymbol{w}) \in [0; 1]$. We want to transform the probability to remove the range restrictions, as $\boldsymbol{x_i}\boldsymbol{w}$ can take any real value.

Using odds

Odds

p – probability of an event occurring

$1 - p$ – probability of the event not occurring

The odds for event i are then defined as

$$\text{odds}_i = \frac{p_i}{1 - p_i}$$

Taking the log of the odds removes the floor and ceiling restrictions.

$$\log\left(\frac{p_i}{1 - p_i}\right) = \sum_j w_j x_{ij} = \boldsymbol{x_i w}$$

This way we map the probabilities from the $[0; 1]$ range to the entire number line.

Logistic function

Logistic Regression Model

$$\log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{x_i w}$$

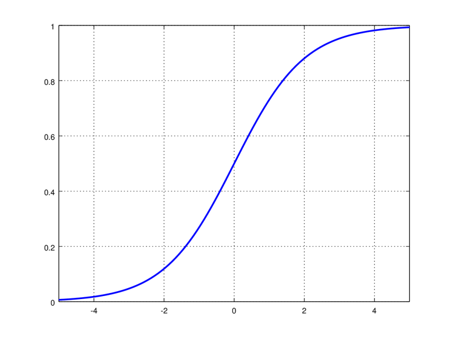
$$\frac{p_i}{1-p_i} = e^{\boldsymbol{x_i w}}$$

$$p_i = \frac{e^{\boldsymbol{x_i w}}}{1+e^{\boldsymbol{x_i w}}} = \frac{1}{1+e^{-\boldsymbol{x_i w}}}$$

$$p(y_i = 1|\boldsymbol{x_i};\boldsymbol{w}) = \frac{1}{1+e^{-\boldsymbol{x_i w}}}$$

$$p(y_i = 0|\boldsymbol{x_i};\boldsymbol{w}) = 1 - \frac{1}{1+e^{-\boldsymbol{x_i w}}}$$

$$p(y_i|\boldsymbol{x_i};\boldsymbol{w}) = \left(\frac{1}{1+e^{-\boldsymbol{x_i w}}}\right)^{y_i}\left(1 - \frac{1}{1+e^{-\boldsymbol{x_i w}}}\right)^{1-y_i}$$



Standard logistic sigmoid function

Estimation

$$\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}^\mathsf{T}, \text{ where } \boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{iD}\}$$

$$\boldsymbol{y} = \{y_1, y_2, \ldots, y_N\}^\mathsf{T}, \text{ where } y_i \in \{0, 1\}$$

$$\boldsymbol{w} = \{w_1, w_2, \ldots, w_D\}^\mathsf{T}$$

Maximum Likelihood Estimation (MLE)

1. **Step 1** – Specify the joint density function

$$p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{w}) = \prod_{i=1}^{N} \left(\frac{1}{1+e^{-\boldsymbol{x_i w}}}\right)^{y_i} \left(1 - \frac{1}{1+e^{-\boldsymbol{x_i w}}}\right)^{1-y_i}$$

2. **Step 2** – Express this is a function of $\boldsymbol{w}$, where $\boldsymbol{X}$ and $\boldsymbol{y}$ are fixed parameters – $L(\boldsymbol{w}) = p(\boldsymbol{y}|\boldsymbol{X};\boldsymbol{w})$
3. **Step 3** – Maximize $L(\boldsymbol{w})$

   $\boldsymbol{w}_{\text{MLE}} = \text{argmax}_{\boldsymbol{w}} L(\boldsymbol{w})$

Likelihood Maximization

$$L(\boldsymbol{w}) = \prod_{i=1}^{N} \left(\frac{1}{1 + e^{-\boldsymbol{x}_i \boldsymbol{w}}} \right)^{y_i} \left(1 - \frac{1}{1 + e^{-\boldsymbol{x}_i \boldsymbol{w}}} \right)^{1-y_i}$$

We can simplify $L(\boldsymbol{w})$ by taking its log and then differentiate to get the gradient.

$$\ell(\boldsymbol{w}) = \sum_{i=1}^{N} \left[y_i \text{log} \left(\frac{1}{1 + e^{-\boldsymbol{x}_i \boldsymbol{w}}} \right) + (1 - y_i) \left(1 + \frac{1}{1 + e^{-\boldsymbol{x}_i \boldsymbol{w}}} \right) \right]$$

$$\nabla_w \ell(\boldsymbol{w}) = \nabla_w \sum_{i=1}^{N} \left[y_i \text{log} \left(\frac{1}{1 + e^{-\boldsymbol{x}_i \boldsymbol{w}}} \right) + (1 - y_i) \left(1 + \frac{1}{1 + e^{-\boldsymbol{x}_i \boldsymbol{w}}} \right) \right]$$

Derivative of the sigmoid

Let $\sigma(x) = \frac{1}{1+e^{-x}}$

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{d}{dx}(1+e^{-x})^{-1} =$$
$$-(1+e^{-x})^{-2}(-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2} =$$
$$\frac{1}{(1+e^{-x})}\frac{-e^{-x}}{(1+e^{-x})} = \frac{1}{(1+e^{-x})}\frac{(1+e^{-x})-1}{(1+e^{-x})} =$$
$$\frac{1}{(1+e^{-x})}\left(1-\frac{1}{(1+e^{-x})}\right) = \sigma(x)(1-\sigma(x))$$

Derivative of the log likelihood

$$\nabla_w \ell(\boldsymbol{w}) = \nabla_w \sum_{i=1}^{N} \left[y_i \mathsf{log} \left(\frac{1}{1+e^{-\boldsymbol{x_i}\boldsymbol{w}}} \right) + (1-y_i) \left(1 + \frac{1}{1+e^{-\boldsymbol{x_i}\boldsymbol{w}}} \right) \right] =$$

$$\nabla_w \sum_{i=1}^{N} \left[y_i \mathsf{log}(\sigma(\boldsymbol{x_i}\boldsymbol{w})) + (1-y_i) \mathsf{log}(1-\sigma(\boldsymbol{x_i}\boldsymbol{w})) \right] =$$

$$\sum_{i=1}^{N} \left(y_i \frac{1}{\sigma(\boldsymbol{x_i}\boldsymbol{w})} \sigma(\boldsymbol{x_i}\boldsymbol{w})(1-\sigma(\boldsymbol{x_i}\boldsymbol{w})\boldsymbol{x_i} + (1-y_i) \frac{1}{\sigma(\boldsymbol{x_i}\boldsymbol{w})} (-1)\sigma(\boldsymbol{x_i}\boldsymbol{w})\boldsymbol{x_i} \right) =$$

$$\sum_{i=1}^{N} \Big(y_i(1-\sigma(\boldsymbol{x_i}\boldsymbol{w}))\boldsymbol{x_i} + (1-y_i)(-1)\sigma(\boldsymbol{x_i}\boldsymbol{w}))\boldsymbol{x_i} \Big) =$$

$$\sum_{i=1}^{N} \Big(y_i\boldsymbol{x_i} - y_i\sigma(\boldsymbol{x_i}\boldsymbol{w})\boldsymbol{x_i} - \sigma(\boldsymbol{x_i}\boldsymbol{w})\boldsymbol{x_i} + y_i\sigma(\boldsymbol{x_i}\boldsymbol{w})\boldsymbol{x_i} \Big) =$$

$$\sum_{i=1}^{N} (y_i\boldsymbol{x_i} - \sigma(\boldsymbol{x_i}\boldsymbol{w})\boldsymbol{x_i}) = \sum_{i=1}^{N} \left(y_i - \frac{1}{(1+e^{-\boldsymbol{x_i}\boldsymbol{w}})} \right) \boldsymbol{x_i}$$

Likelihood Maximization

We can now use gradient ascent to maximize $\ell(\boldsymbol{w})$

The update rule will be:

repeat until convergence $\{$

$$w_j := w_j + \alpha \sum_{i=1}^{N} \Big(y_i - \frac{1}{(1 + e^{-\boldsymbol{x_i w}})} \Big) x_{ij}$$

$\}$

or using matrix notation

repeat until convergence $\{$

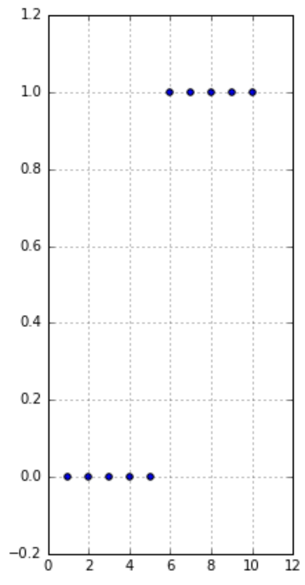
$$\boldsymbol{w} := \boldsymbol{w} + \alpha \boldsymbol{X}^{\mathsf{T}} \left(\boldsymbol{y} - \frac{1}{1 + e^{-\boldsymbol{X}\boldsymbol{w}}} \right)$$

$\}$

Example

Simple example

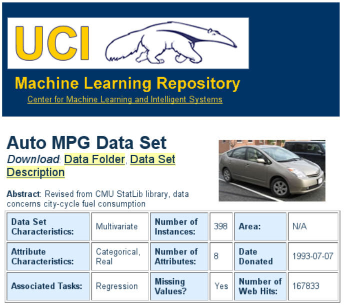
$$\boldsymbol{X} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}^\mathsf{T}$$
$$\boldsymbol{y} = \{0, 0, 0, 0, 0, 1, 1, 1, 1, 1\}^\mathsf{T}$$



Data Set Example

UCI Machine Learning Repository –
`archive.ics.uci.edu/ml`

- Great resource for Machine Learning data sets
- Over 330 freely available sets
- Auto MPG Data Set
  - Fuel consumption in MPG
  - Attributes: mpg, cylinders, displacement, horsepower, weight, acceleration etc.



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Auto MPG Data Set
Download: Data Folder, Data Set Description

Abstract: Revised from CMU StatLib library, data concerns city-cycle fuel consumption

Data Set Characteristics:	Multivariate	**Number of Instances:**	398	**Area:**	N/A
Attribute Characteristics:	Categorical, Real	**Number of Attributes:**	8	**Date Donated**	1993-07-07
Associated Tasks:	Regression	**Missing Values?**	Yes	**Number of Web Hits:**	167833

Non linear decision boundary

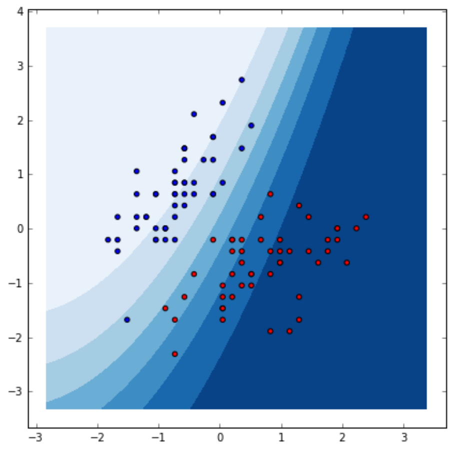
Polynomial predictors

- Relationship is modelled using a k^{th} degree polynomial
- The hypothesis is then

$$\hat{y}(x_i) = \left(\frac{1}{1 + e^{-\boldsymbol{x_i w}}} \right)$$

where

$$\boldsymbol{x_i w} = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_k x_i^k$$



Final remarks

No analytical solution

Assumptions

- Not as strict as Linear Regression (e.g. no assumption on homoscedasticity, no linear relationship between dependent and independent variables, residual do not need to be normally distributed etc.)
- There are still certain assumptions
  - Linear relationship between the $logit$ of the independent variables and the dependent variable
  - Binary dependent variable
  - Independent error terms
  - The sample is sufficiently large – check [Hsi89]

References I

- Sami Abu-El-Haija, *Derivation of logistic regression*, www.haija.org/derivation_logistic_regression.pdf, Accessed: 20-03-2017.
- Christopher M. Bishop, *Pattern recognition and machine learning (information science and statistics)*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
- Peter Flach, *Machine learning: The art and science of algorithms that make sense of data*, Cambridge University Press, New York, NY, USA, 2012.
- F. Y. Hsieh, *Sample size tables for logistic regression*, Statistics in Medicine **8** (1989), no. 7, 795–802.

Q&A

