Classification & Regression

Course Topics

Preliminaries

Data Understanding

Data Preprocessing

Clustering & Association

Classification & Regression

Validation & Interpretation

Advanced Topics

Done!

Definitions

- Unsupervised Learning:
  - Given items X, automatically discover the structure, representations, etc.
  - Ex.: Clustering
- Supervised Learning:
  - Given the value of an input vector X and $c(x)$, predict c on future unseen x's.
  - Ex.: Regression, classification

Supervised Learning

- Regression
  - Given the value of an input X, the output Y belongs to the set of real values R. Goal is to predict output accurately for new input.
- Classification
  - The predictions or outputs, $c(x)$ are categorical while x can take any set of values (real or categorical). The goal is select correct class for a new instance.
- Time series prediction
  - Data is in the form of a moving time series. The goal is to perform classification/regression on future time series based on data known so far.

Regression

- Predictive technique where the target variable to be estimated is continuous.
  - Applications:
    - Predicting the stock market index based on economic factors
    - Forecasting precipitation based on characteristics of the jet stream
    - Projecting a company's revenue based on the amount spent for advertisement

Regression

- Let D denote a dataset containing N observations,
$$D = \{(x_i, y_i) | i = 1,2, \dots, N\}$$
  - Each x_i corresponds to the set of attributes of the i-th observation. These are called **explanatory variables** and can be discrete or continuous.
  - y_i corresponds to the **target variable**.

Definition. Regression is the task of learning a target function f that maps each attribute set x into a continuous-valued output y.

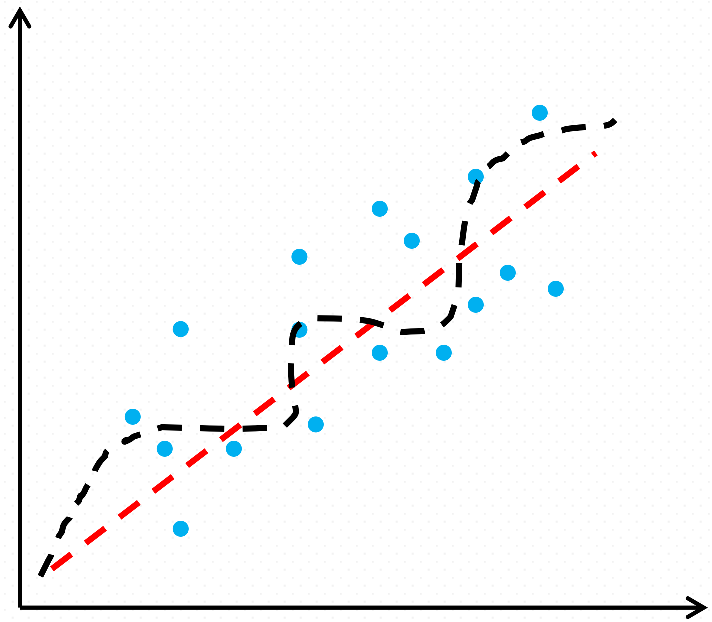
Error function

- The goal of linear regression is to find a target function that can minimize the error, which can be measured as the sum of absolute or squared error.

$$Absolute\ Error = \sum_i |y_i - f(x_i)|$$

$$Squared\ Error = \sum_i \left(y_i - f(x_i)\right)^2$$

Simple Linear Regression



Given a set of points (x_i, y_i) on a scatterplot

Find the best-fit line $f(x_i) = \boldsymbol{w_0} + \boldsymbol{w_1} x_i$

Such that SSE $= \sum_i (y_i - f(x_i))^2$ is minimized

Simple Linear Regression

- To find the regression parameters w_0 and w_1, we apply the **method of least squares**, which attempts to minimize the SSE

$$SSE = \sum_{i=1}^{N} [y_i - f(x_i)]^2 = \sum_{i=1}^{N} [y_i - w_0 - w_1 x_i]^2$$

Simple Linear Regression

- This optimization problem can be solved by taking the partial derivatives of E with respect to w_0 and w_1, setting them to 0 and solving the system of linear equations.

$$\frac{\partial E}{\partial w_0} = -2\sum_{i=1}^{N}[y_i - w_1 x_i - w_0] = 0$$

$$\frac{\partial E}{\partial w_1} = -2\sum_{i=1}^{N}[y_i - w_1 x_i - w_0]x_i = 0$$

Simple Linear Regression

- The previous equations can be summarized by the **normal equation:**

$$\begin{pmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{pmatrix}$$

Example

- Consider the set of 10 points:

x	1	2	3	4	4	5	5	6	6	7
y	7	8	9	8	9	11	10	13	14	13

$$\sum_i x_i = 43 \qquad \sum_i x_i^2 = 217$$

$$\sum_i y_i = 102 \qquad \sum_i y_i^2 = 1094$$

$$\sum_i x_i y_i = 476$$

Example

x	1	2	3	4	4	5	5	6	6	7
y	7	8	9	8	9	11	10	13	14	13

$$\sum_i x_i = 43 \quad \sum_i x_i^2 = 217 \quad \sum_i y_i = 102 \quad \sum_i y_i^2 = 1094 \quad \sum_i x_i y_i = 476$$

$$\begin{pmatrix} 10 & 43 \\ 43 & 217 \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 102 \\ 476 \end{pmatrix}$$

$$\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 10 & 43 \\ 34 & 217 \end{pmatrix}^{-1} \begin{pmatrix} 102 \\ 476 \end{pmatrix}$$

$$\begin{pmatrix} w_0 \\ w_1 \end{pmatrix} = \begin{pmatrix} 5.19 \\ 1.17 \end{pmatrix}$$

$$\boldsymbol{f(x_i) = 1.17 x_i + 5.19}$$

Example

x	1	2	3	4	4	5	5	6	6	7
y	7	8	9	8	9	11	10	13	14	13

- A general solution to the normal equations can be expressed as

$$\boldsymbol{w_0 = \bar{y} - w_1\bar{x}}$$
$$\boldsymbol{w_1 = \frac{\sigma_{xy}}{\sigma_{xx}}}$$

Where $\bar{x}$ and $\bar{y}$ are the average values of x and y and:

$$\sigma_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$
$$\sigma_{xx} = \sum_i (x_i - \bar{x})^2$$
$$\sigma_{yy} = \sum_i (y_i - \bar{y})^2$$

Example

x	1	2	3	4	4	5	5	6	6	7
y	7	8	9	8	9	11	10	13	14	13

- A linear model that results in the minimum squared error is then:

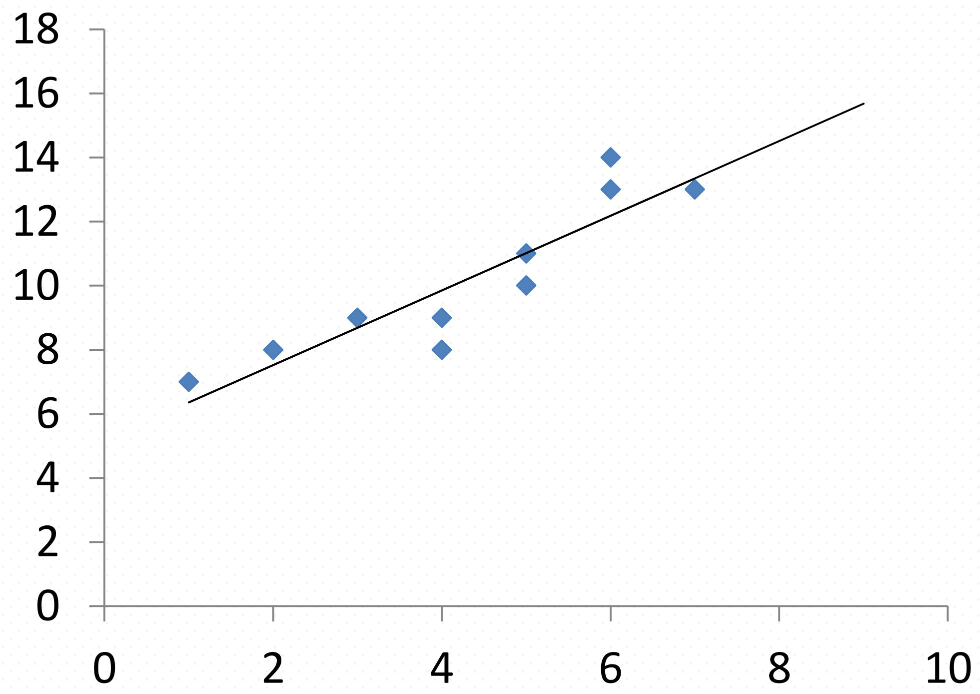
$$\boldsymbol{f(x) = \bar{y} + \frac{\sigma_{xy}}{\sigma_{xx}}(x - \bar{x})}$$

- For our example:

$$f(x) = 10.2 + \frac{37.4}{32.1}(x - 4.3)$$
$$f(x) = 10.2 + 1.17(x - 4.3)$$
$$f(x) = 10.2 + 1.17x - 5.01$$
$$\boldsymbol{f(x) = 1.17x + 5.19}$$

Example

x	1	2	3	4	4	5	5	6	6	7
y	7	8	9	8	9	11	10	13	14	13



Evaluating goodness of fit

- In order to measure how well data points fit to our line, we use a method called **R squared (R^2)**
- This value ranges from 0 to 1. It is close to 1 if most variability observed in the target variable can be explained by the regression model

Evaluating goodness of fit

$$R^2 = \frac{SSM}{SST} = \frac{\sum_i [f(x_i) - \bar{y}]^2}{\sum_i [y - \bar{y}]^2}$$

$$R^2 = \frac{\sigma_{xy}^2}{\sigma_{xx}\sigma_{yy}}$$

- As we add more explanatory variables, R^2 increases, so it is typically adjusted as:

$$Adjusted\ R^2 = 1 - \left(\frac{N-1}{N-d}\right)(1 - R^2),$$

Where N is the number of data points and d+1 is the number of parameters of the regression model.

Logistic Regression

- Useful when the target is binary
- Logistic regression or logit regression is a type of probabilistic statistical classification model
- It measures the relationship between the dependent (target) binary variable and the independent explanatory variables

Logistic Regression

- In summary:
  - We have a binary target variable Y, and we want to model the conditional probability $P(Y = 1|X = x)$ as a function $p(x)$ of the explanatory variables x.
  - Any unknown parameters (recall w_0 and w_1) are estimated by maximum likelihood.
  - Can we use linear regression?

Logistic Regression

- Idea 1: Let $p(x)$ be a linear function
  - W are estimating a probability, which must be between 0 and 1
  - Linear functions are unbounded, so this approach doesn't work
- Better idea: Set the odds ratio to a linear function:

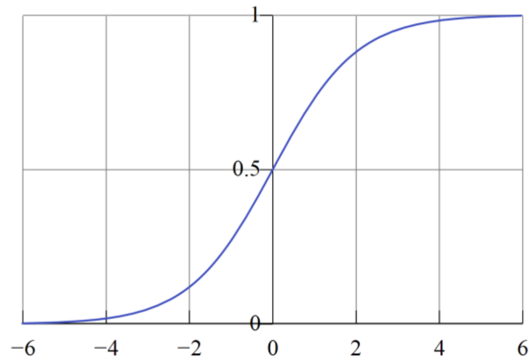
$$\log(odds) = logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

Solving for p:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{\mathbf{1}}{\mathbf{1}+\boldsymbol{e}^{-(\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{x})}}$$

  - This is called the logistic (logit) function and it assumes values [0,1]

Logistic Curve



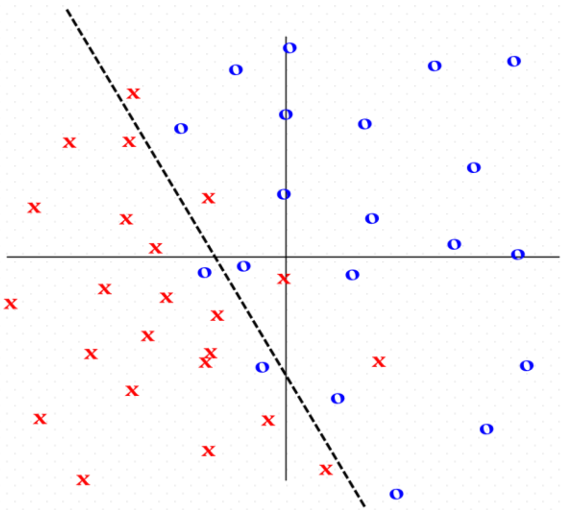
- A sigmoid function that assumes values in the range [0,1]

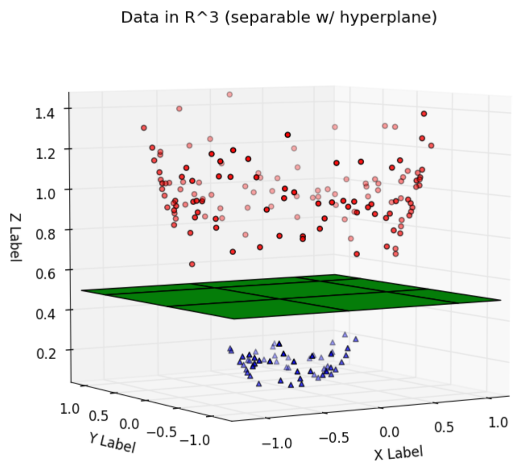
$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

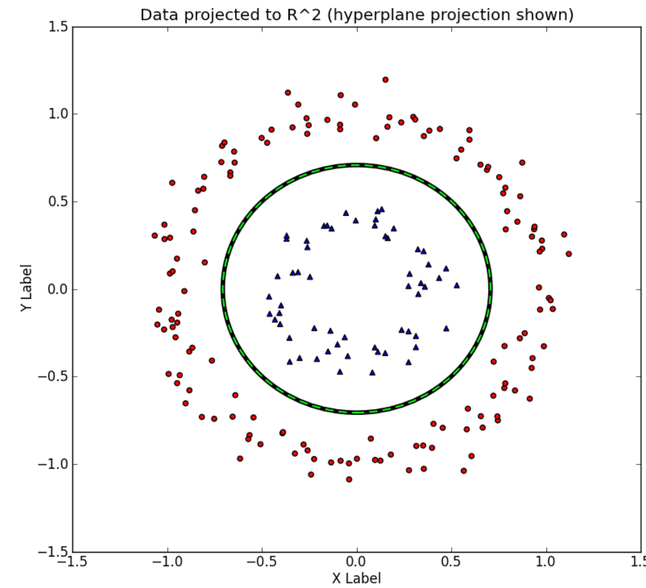
Logistic Regression

- To minimize misclassification rates, we predict:
  - $Y = 1$ when $p(x) \geq 0.5$ and $Y = 0$ when $p(x) < 0.5$
  - So $Y = 1$ when $\beta_0 + \beta_1 x$ is non-negative and 0 otherwise
- Logistic regression gives us a **linear classifier** where the decision boundary separating the two classes is the solution of $\beta_0 + \beta_1 x = 0$
  - A point if we have 1 explanatory variable
  - A line if we have 2
  - A plane if we have 3
  - A disaster if we have more than that

Decision Boundaries







Logistic Regression

- The parameters $\beta_0, \beta_1, \ldots$ are estimated using a technique called Maximum likelihood estimation (MLE)
  - Unlike the least squares methods used for Linear regression, finding a closed form for the coefficients using MLE is not possible. Instead, an iterative process (e.g., Newton's method) is used.
  - This process begins with a tentative solution, revises it slightly to see if it can be improved, and repeats this revision until improvement is minute, at which point the process is said to have converged.

Logistic Regression

- Goodness of fit for logistic regression can't be measured using R^2. Methods used in this case include:
  - Hosmer–Lemeshow test
  - Binary classification performance evaluation
  - Deviance and likelihood ratio tests

Now lets see some regressioning!