GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

CLASSIFICATION, LOGISTIC REGRESSION

AGENDA

- Classification Problems
- Logistic Regression
- Evaluating Logistic Regression Models

CLASSIFICATION

CLASSIFICATION VS. CLUSTERING?

Classification:

Clustering:

CLASSIFICATION VS. CLUSTERING?

Classification: Supervised

Clustering: Unsupervised

Classification

- Predict the value of a categorical response variable

CLASSIFICATION

- Predict the value of a categorical response variable
- Approximate a function that maps an observation to its associated class or label

BINARY CLASSIFICATION

Assign an instance to one of two classes:

- predict whether an image depicts a cat or a dog
- predict whether a tumor is malignant or benign

Multi-class Classification

Assign an instance to one of more than two classes:

- predict if an image depicts a cat, dog or bird
- predict if a news article should be included in the sports, politics, leisures, or business sections

Multi-label Classification

Assign instances to one or more of more than two classes:

- predict which of the following tags pertain to a StackOverflow question: Java, Guava, C#, dependency injection, REST, Jersey, Jackson
- predict if a news article should be categorized into multiple classes such as sports and/or current events and/or entertainment

PERFORMANCE MEASURES

How do we measure how correct a classification model is?

PERFORMANCE MEASURES

Predict whether tumors are malignant or benign:

- ***Accuracy*****:** fraction of instances that are classified correctly
  - does not differentiate between malignant tumors that were classified as being benign, and benign tumors that were classified as being malignant.
- In some problems, the costs associated with all types of errors may be the same
- In this problem, failing to identify malignant tumors is likely more severe than failing to identify benign tumors as malignant

PERFORMANCE MEASURES

- **True positive**: correctly classifying a malignant tumor

PERFORMANCE MEASURES

- **True positive**: correctly classifying a malignant tumor
- **True negative**: correctly classifying a benign tumor

PERFORMANCE MEASURES

- **True positive**: correctly classifying a malignant tumor
- **True negative**: correctly classifying a benign tumor
- **False positive**: a benign tumor that is incorrectly classifier as being malignant

PERFORMANCE MEASURES

- **True positive**: correctly classifying a malignant tumor
- **True negative**: correctly classifying a benign tumor
- **False positive**: a benign tumor that is incorrectly classifier as being malignant
- **False negative**: a malignant tumor that is incorrectly classifier as being benign

CONFUSION MATRIX

	Prediction		
Actual		1	0
	1	**TP**	**FP**
	0	**FN**	**TN**

PERFORMANCE MEASURES

- ***Accuracy*** is the fraction of instances that were classified correctly

$$\mathrm{ACC} = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})}$$

PERFORMANCE MEASURES

- ***Precision*** is the fraction of the tumors that were predicted to be malignant that are actually malignant.

$$P = TP / (TP + FP)$$

PERFORMANCE MEASURES

- ***Recall*** (or True Positive Rate) is the fraction of malignant tumors that the system identified.

$$R = TP / (TP + FN)$$

PERFORMANCE MEASURES

- ***Fall-out*** *or false positive rate* ***(FPR):***

$$\mathrm{FPR} = \mathrm{FP} / (\mathrm{FP} + \mathrm{TN})$$

LOGISTIC REGRESSION

Logistic Regression

Q: What is logistic regression?

LOGISTIC REGRESSION

Q: What is logistic regression?

A: A generalization of the linear regression model to *classification* problems.

Logistic Regression

In linear regression, we used a set of covariates to predict the value of a (continuous) outcome variable.

Logistic Regression

In linear regression, we used a set of covariates to predict the value of a (continuous) outcome variable.

In logistic regression, we use a set of covariates to predict probabilities of (binary) class membership.

Logistic Regression

In linear regression, we used a set of covariates to predict the value of a (continuous) outcome variable.

In logistic regression, we use a set of covariates to predict probabilities of (binary) class membership.

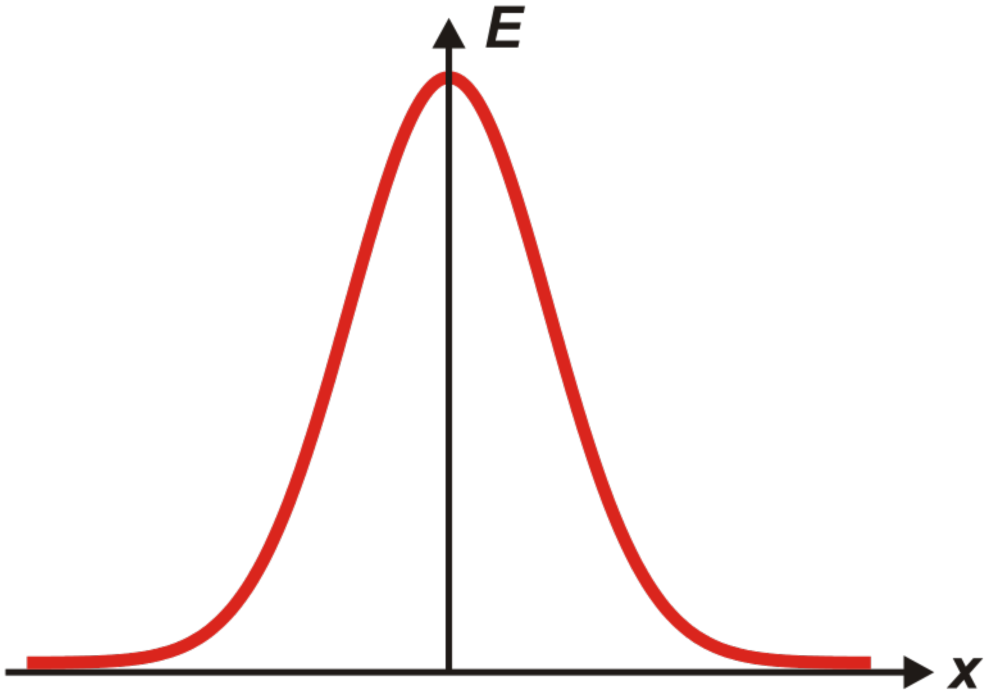
These probabilities are then mapped to class labels, thus solving the classification problem.

RESPONSE VARIABLE DISTRIBUTIONS

- Ordinary linear regression assumes that the response variable is normally distributed.

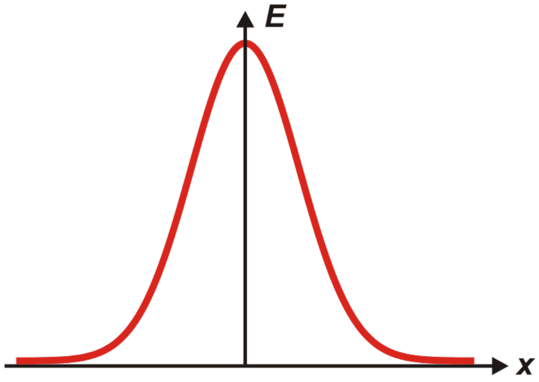
NORMAL DISTRIBUTION

- Gaussian distribution, or bell curve



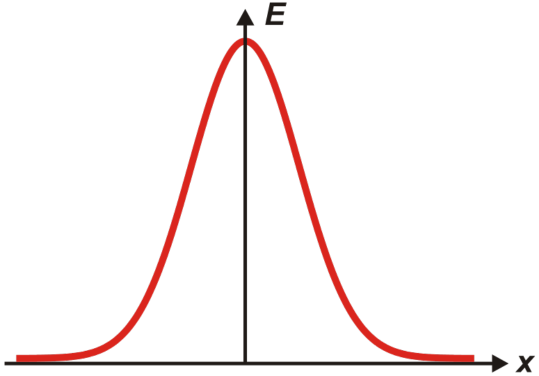
NORMAL DISTRIBUTION

- Describes the probability that an observation will have a value between any two real numbers.



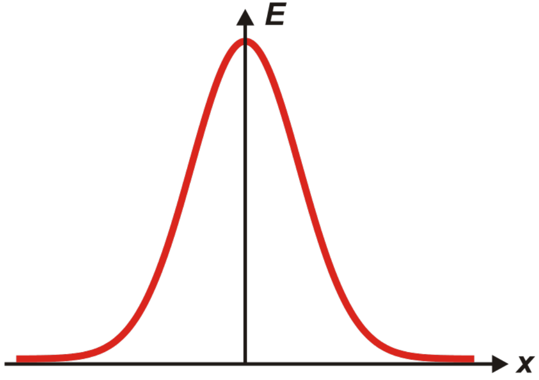
NORMAL DISTRIBUTION

- Describes the probability that an observation will have a value between any two real numbers.
- Normally distributed data is symmetrical.



NORMAL DISTRIBUTION

- Describes the probability that an observation will have a value between any two real numbers.
- Normally distributed data is symmetrical.
- The mean, median, and mode of normally distributed data are equal.



NORMAL DISTRIBUTION

- Describes the probability that an observation will have a value between any two real numbers.
- Normally distributed data is symmetrical.
- The mean, median, and mode of normally distributed data are equal.
- Many natural phenomena approximately follow normal distributions. E.g., the heights of people are normally distributed.

RESPONSE VARIABLE DISTRIBUTIONS

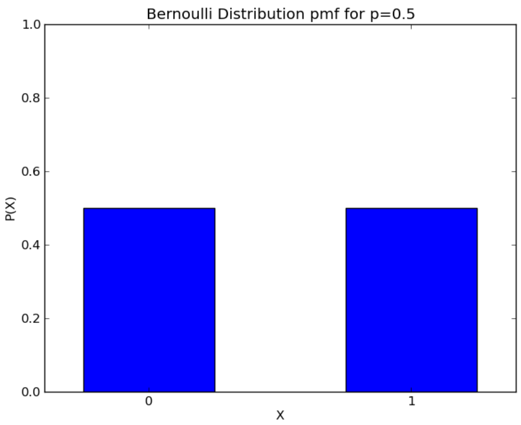
- In some problems the response variable is not normally distributed.

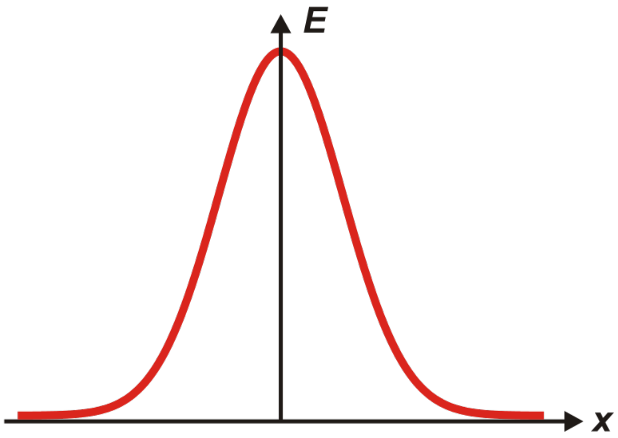
RESPONSE VARIABLE DISTRIBUTIONS

- In some problems the response variable is not normally distributed.
- A coin toss can result in two outcomes: heads or tails.

RESPONSE VARIABLE DISTRIBUTIONS

- In some problems the response variable is not normally distributed.
- A coin toss can result in two outcomes: heads or tails.





RESPONSE VARIABLE DISTRIBUTIONS

- In some problems the response variable is not normally distributed.
- A coin toss can result in two outcomes: heads or tails.
- The **Bernoulli distribution** describes the probability distribution of a random variable that can take the positive case with probability ***P*** or the negative case with probability ***1 - P***.

RESPONSE VARIABLE DISTRIBUTIONS

- Linear regression assumes that a constant change in the value of an explanatory variable results in a constant change in the value of the response variable.

RESPONSE VARIABLE DISTRIBUTIONS

- Linear regression assumes that a constant change in the value of an explanatory variable results in a constant change in the value of the response variable.

- In Logistic regression, the response variable represents a probability that must be constrained to the range {0, 1}.

LINEAR REGRESSION FOR CLASSIFICATION?

Suppose we encode a response variable ***Y*** as {0=No, 1=Yes}

- Can we simply perform a linear regression of ***Y*** on ***X*** and classify as "*Yes*" if ***Y > 0.5*** ?

LINEAR REGRESSION FOR CLASSIFICATION?

Suppose we encode a response variable ***Y*** as {0=No, 1=Yes}

- Can we perform a linear regression of ***Y*** on ***X*** and simply classify as "*Yes*" if ***Y > 0.5*** ?
    - In the case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to ***linear discriminant analysis***

LINEAR REGRESSION FOR CLASSIFICATION?

Suppose we encode a response variable ***Y*** as {0=No, 1=Yes}

- Can we perform a linear regression of ***Y*** on ***X*** and simply classify as "*Yes*" if ***Y > 0.5*** ?
    - In the case of a binary outcome, linear regression does a decent job as a classifier, and is equivalent to ***linear discriminant analysis***
    - However, linear regression might produce probabilities ***P < 0*** or ***P > 1***...

LINEAR REGRESSION FOR CLASSIFICATION?

Suppose we have a response variable with three possible values. A patient arrives at the emergency room, and we must classify them according to their symptoms.

$$\mathtt{Y} = \begin{Bmatrix} \mathtt{1,\ stroke} \\ \mathtt{2,\ drug\ OD} \\ \mathtt{3,\ heart\ attack} \end{Bmatrix}$$

This encoding implies that the difference between stroke and drug overdose is the same as between drug overdose and heart attack

LINK FUNCTIONS

- ***Generalized linear models*** relate a linear combination of the explanatory variables and model parameters to the response variable using a ***link function***.

LINK FUNCTIONS

- ***Generalized linear models*** relate a linear combination of the explanatory variables and model parameters to the response variable using a ***link function***.

- Ordinary linear regression is a special case of the generalized linear model that relates a linear combination of the explanatory variables to a normally-distributed response variable using the ***identity link function***.

LINK FUNCTIONS

- ***Generalized linear models*** relate a linear combination of the explanatory variables and model parameters to the response variable using a ***link function***.
- Ordinary linear regression is a special case of the generalized linear model that relates a linear combination of the explanatory variables to a normally-distributed response variable using the ***identity link function***.
- We can use a different link function to relate a linear combination of the explanatory variables to response variable that is not normally-distributed.

LINK FUNCTIONS

- Logistic regression regresses the probability that an instance is the positive case onto the explanatory variables.

LINK FUNCTIONS

- Logistic regression regresses the probability that an instance is the positive case onto the explanatory variables.
- Models the conditional probability $\boldsymbol{P(Y = 1 \mid X)}$

LINK FUNCTIONS

- Logistic regression regresses the probability that an instance is the positive case onto the explanatory variables.
- Models the conditional probability ***P(Y = 1 | X)***
- If the response variable is equal to or exceeds a discrimination threshold the positive class is predicted; otherwise, the negative class is predicted.

LINK FUNCTIONS

- Logistic regression regresses the probability that an instance is the positive case onto the explanatory variables.
- Models the conditional probability ***P(Y = 1 | X)***
- If the response variable is equal to or exceeds a discrimination threshold the positive class is predicted; otherwise, the negative class is predicted.
- The response variable is modeled as a function of a linear combination of the explanatory variables using the **logistic function**.

THE LOGISTIC FUNCTION

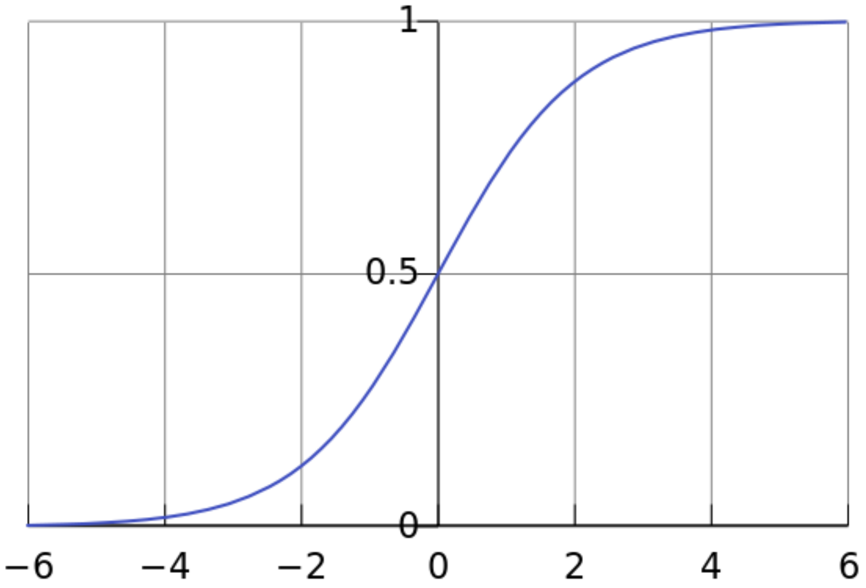
- The logistic function always returns a value between zero and one.

$$F(t) = \frac{1}{1 + e^{-t}}$$

THE LOGISTIC FUNCTION

- The logistic function always returns a value between zero and one.

$$F(t) = \frac{1}{1 + e^{-t}}$$



THE LOGISTIC FUNCTION

$$F(t) = \frac{1}{1 + e^{-t}}$$

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta x)}}$$

THE LOGISTIC FUNCTION

- The **logit function** is the inverse of the logistic function. It links back to a linear combination of the explanatory variables so that the parameter values can be solved.

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta x)}}$$

$$g(x) = ln \frac{F(x)}{1 - F(x)} = \beta_0 + \beta x$$

PRACTICE