

INTRO to DATA SCIENCE

LECTURE 5: PROBABILITY & NAIVE BAYESIAN CLASSIFICATION

LAST TIME:

- LINEAR REGRESSION**
- REGULARIZATION**

QUESTIONS?

I. REVIEW OF REGULARIZATION

II. PROBABILITY

III. LOGISTIC REGRESSION

These regularization problems can also be expressed as:

OLS: $\min(\|y - x\beta\|^2)$

L1 regularization: $\min(\|y - x\beta\|^2 + \lambda\|x\|)$

L2 regularization: $\min(\|y - x\beta\|^2 + \lambda\|x\|^2)$

We are no longer just minimizing error but also an additional term.

These regularization problems can also be expressed as:

L1 regularization: $\min(\|y - x\beta\|^2 + \lambda\|x\|)$

When do we use L1?

These regularization problems can also be expressed as:

L1 regularization: $\min(\|y - x\beta\|^2 + \lambda\|x\|)$

Common case: Text Classification

$X = [\text{animal} = 1, \dots, \text{carnival} = 0, \dots, \text{xylophone} = 0, \dots, \text{zebra} = 0]$

$Y = \text{Topic}$ or $Y = \text{Important/Not Important}$ or $Y = \text{Positive/Negative}$

I. INTRO TO PROBABILITY

Q: What is a **probability**?

Q: What is a **probability**?

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

Q: What is a **probability**?

A: A number between 0 and 1 that characterizes the likelihood that some event will occur.

The probability of event A is denoted $P(A)$.

Q: What is the set of all possible events called?

Q: What is the set of all possible events called?

A: This set is called the **sample space** Ω . Event A is a member of the sample space, as is every other event.

Q: What is the set of all possible events called?

A: This set is called the **sample space** Ω . Event A is a member of the sample space, as is every other event.

The probability of the sample space $P(\Omega)$ is 1.

Q: What is a probability distribution?

A: A function that assigns probability to each event in the sample space.

Q: What is a probability distribution?

A: A function that assigns probability to each event in the sample space.

A distribution can be *discrete* or *continuous*

Ex:

Discrete – Uniform distribution

$$X \sim \{1, \dots, N\} \quad - \quad P(X = x) = 1/N$$

Q: What is a probability distribution?

A: A function that assigns probability to each event in the sample space.

A distribution can be *discrete* or *continuous*

Ex:

Continuous – Normal distribution – $N(u, \sigma)$

$$X \sim N(0, 1) \quad - \quad P(X = x) = 0$$

Q: What is expected value?

A: It is the average value of a random variable – one that represents the most common value

Q: What is expected value?

A: It is the average value of a random variable – one that represents the most common value

For discrete distributions

$$E(X) = \sum x * p(x)$$

For continuous distributions

$$E(X) = \text{integral} (x * p(x))$$

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

- 1) Linda is a bank teller.
- 2) Linda is a bank teller and active in the feminist movement.

Q: Consider two events A & B . How can we characterize the intersection of these events?

Q: Consider two events A & B . How can we characterize the intersection of these events?

A: With the joint probability of A and B , written $P(AB)$.

Q: Suppose event B has occurred. What quantity represents the probability of A **given** this information about B ?

Q: Suppose event B has occurred. What quantity represents the probability of A **given** this information about B ?

A: The intersection of A & B divided by region B .

Q: Suppose event B has occurred. What quantity represents the probability of A **given** this information about B ?

A: The intersection of A & B divided by region B .

Q: Suppose event B has occurred. What quantity represents the probability of A **given** this information about B ?

A: The intersection of A & B divided by region B .

This is called the **conditional probability** of A given B , written $P(A|B) = P(AB) / P(B)$.

Q: Suppose event B has occurred. What quantity represents the probability of A **given** this information about B ?

A: The intersection of A & B divided by region B .

This is called the **conditional probability** of A given B , written $P(A|B) = P(AB) / P(B)$.

Notice, with this we can also write $P(AB) = P(A|B) * P(B)$.

Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

- 1) Linda is a bank teller.
- 2) Linda is a bank teller and active in the feminist movement.

Q: What does it mean for two events to be independent?

Q: What does it mean for two events to be **independent**?

A: Information about one does not affect the probability of the other.

Q: What does it mean for two events to be **independent**?

A: Information about one does not affect the probability of the other.

This can be written as $P(A|B) = P(A)$.

Q: What does it mean for two events to be **independent**?

A: Information about one does not affect the probability of the other.

This can be written as $P(A|B) = P(A)$.

Using the definition of the conditional probability, we can also write:

$$P(A|B) = P(AB) / P(B) = P(A) \rightarrow P(AB) = P(A) * P(B)$$

This result is called **Bayes' theorem**. Here it is again:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

This result is called **Bayes' theorem**. Here it is again:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.

This result is called **Bayes' theorem**. Here it is again:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Some facts:

- This is a simple algebraic relationship using elementary definitions.
- It's interesting because it's kind of a “wormhole” between two different “interpretations” of probability.

This result is called **Bayes' theorem**. Here it is again:

$$P(A|B) = P(B|A) * P(A) / P(B)$$

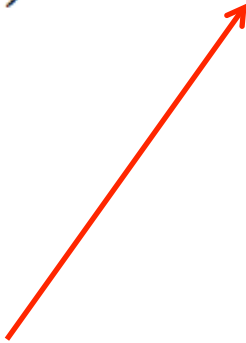
Some facts:

- This is a simple algebraic relationship using elementary definitions.
- It's interesting because it's kind of a “wormhole” between two different “interpretations” of probability.
- It's a very powerful computational tool.

Each term in this relationship has a name, and each plays a distinct role in any probability calculation (including ours).

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

This term is the **likelihood function**. It represents the joint probability of observing features $\{x_i\}$ given that that record belongs to class C .

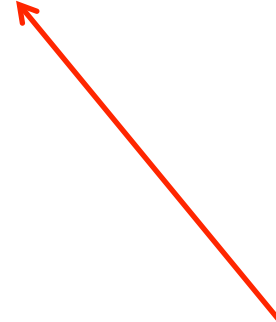
$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


This term is the **likelihood function**. It represents the joint probability of observing features $\{x_i\}$ given that that record belongs to class C .

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

We can observe the value of the likelihood function from the training data.

This term is the **prior probability** of C . It represents the probability of a record belonging to class C before the data is taken into account.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


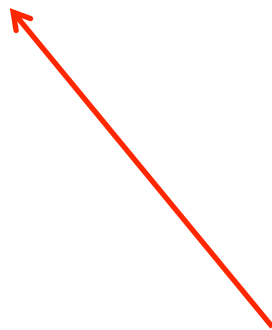
This term is the **prior probability** of C . It represents the probability of a record belonging to class C before the data is taken into account.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

The value of the prior is also observed from the data.

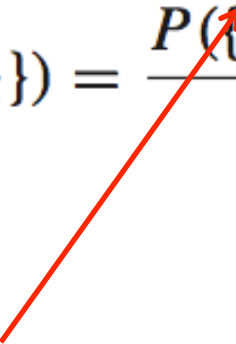
This term is the **normalization constant**. It doesn't depend on C , and is generally ignored until the end of the computation.

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$



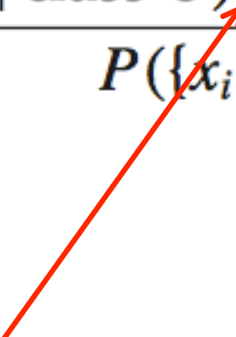
Maximum likelihood estimator (MLE):

What parameters ***maximize*** the likelihood function?

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


Maximum a posteriori estimate (MAP):

What parameters *maximize* the likelihood function **AND** prior?

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


II. LOGISTIC REGRESSION

	<i>continuous</i>	<i>categorical</i>
<i>supervised</i>	regression	classification
<i>unsupervised</i>	dimension reduction	clustering

Q: What is logistic regression?

Q: What is **logistic regression**?

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

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

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

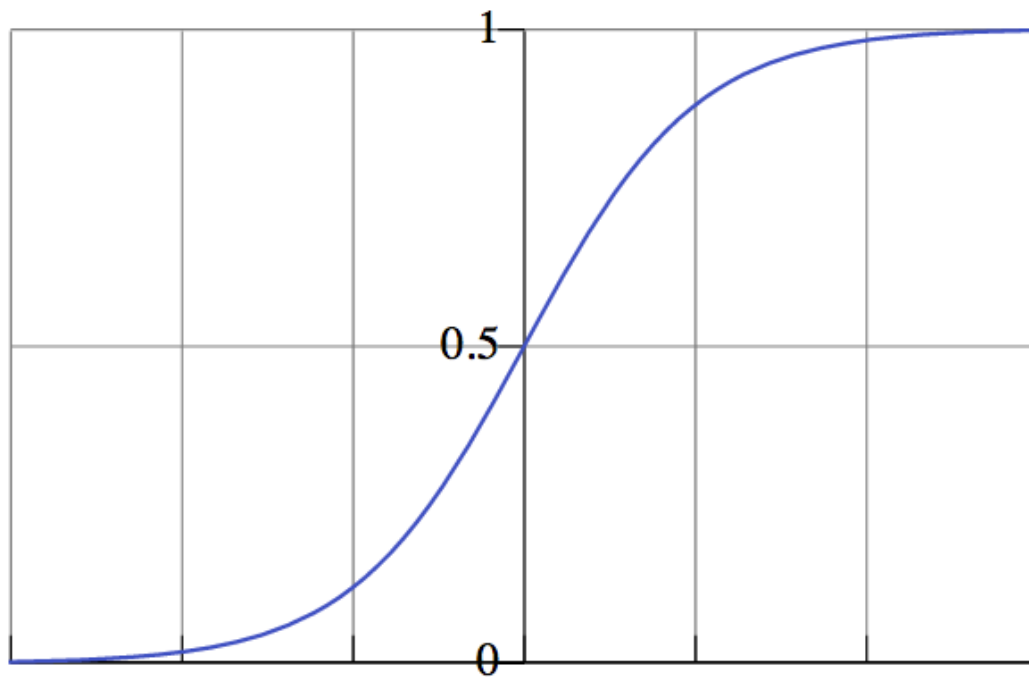
In logistic regression, we use a set of covariates to predict *probabilities* of (binary) class membership.

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.

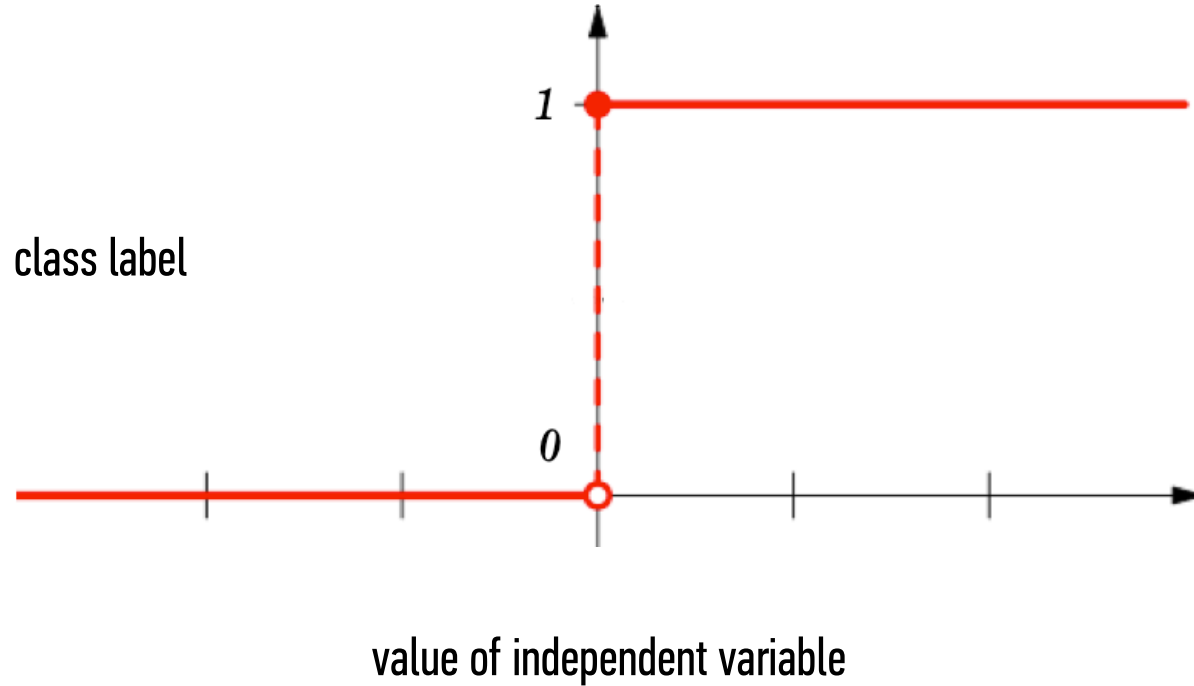
These probabilities are then mapped to *class labels*, thus solving the classification problem.

probability of
belonging to
class



NOTE

Probability predictions
look like this.



NOTE

Probabilities are “snapped” to class labels (eg by thresholding at 50%).

The logistic regression model is an *extension* of the linear regression model, with a couple of important differences.

The logistic regression model is an *extension* of the linear regression model, with a couple of important differences.

The first difference is in the outcome variable.

The logistic regression model is an *extension* of the linear regression model, with a couple of important differences.

The main difference is in the outcome variable.

The key variable in any regression problem is the **response type** of the outcome variable y given the value of the covariate x :

$$E(y|x)$$

The key variable in any regression problem is the **conditional mean** of the outcome variable y given the value of the covariate x :

$$E(y|x)$$

In linear regression, we assume that this conditional mean is a linear function taking values in $(-\infty, +\infty)$:

$$E(y|x) = \alpha + \beta x$$

In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval $[0, 1]$.

In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval $[0, 1]$.

The first step in extending the linear regression model to logistic regression is to map the outcome variable $E(y|x)$ into the unit interval.

In logistic regression, we've seen that the conditional mean of the outcome variable takes values only in the unit interval $[0, 1]$.

The first step in extending the linear regression model to logistic regression is to map the outcome variable $E(y | x)$ into the unit interval.

Q: How do we do this?

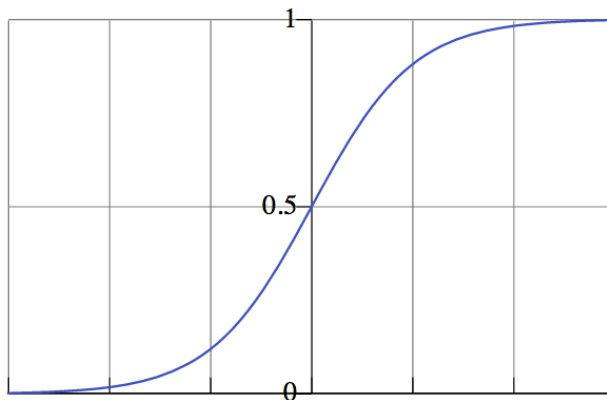
A: By using a transformation called the **logistic function**:

$$E(y|x) = \pi(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

A: By using a transformation called the **logistic function**:

$$E(y|x) = \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

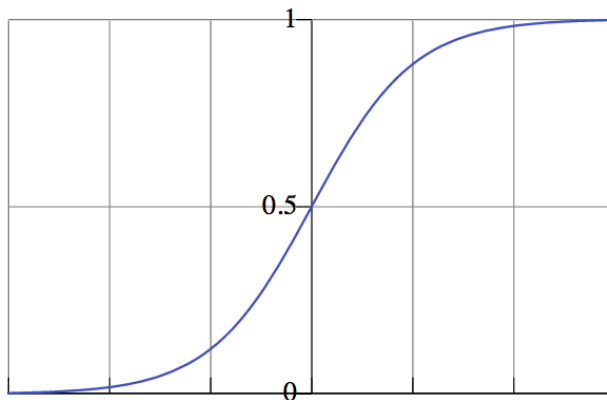
We've already seen what this looks like:



A: By using a transformation called the **logistic function**:

$$E(y|x) = \pi(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$$

We've already seen what this looks like:



NOTE

For any value of x , y is in the interval $[0, 1]$

This is a nonlinear transformation!

The **logit function** is an important transformation of the logistic function. Notice that it returns the linear model!

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

The **logit function** is an important transformation of the logistic function. Notice that it returns the linear model!

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

The logit function is also called the **log-odds function**.

The **logit function** is an important transformation of the logistic function. Notice that it returns the linear model!

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

NOTE

This name hints at its usefulness in interpreting our results.

We will see why shortly.

The logit function is also called the **log-odds function**.