Machine Learning and Intelligent Systems

Linear Models for Classification: Logistic Regression

Maria A. Zuluaga

October 20, 2023

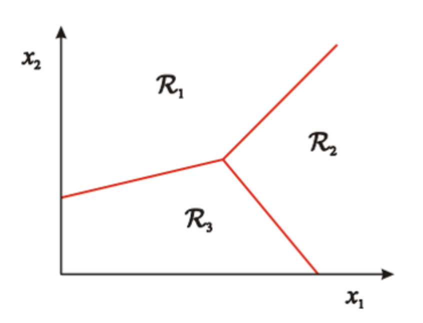
EURECOM - Data Science Department

Table of contents

Quick recap

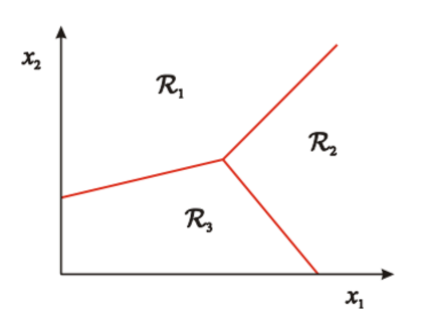
Supervised Learning: Classification

- **Assumption 1:** The target variable (output) y is now binary



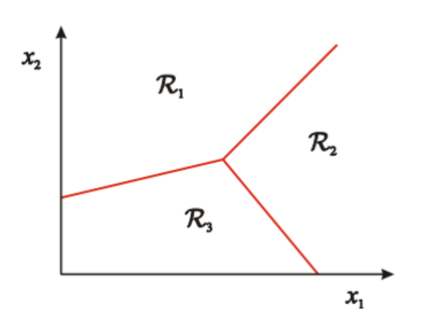
Supervised Learning: Classification

- **Assumption 1:** The target variable (output) y is now binary
- **Assumption 2:** The input data x is separable



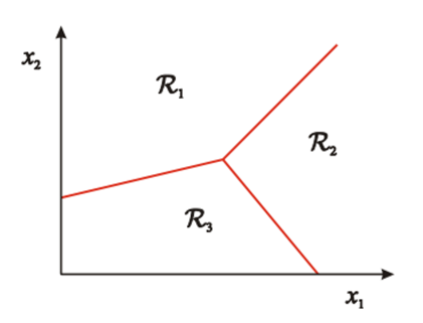
Supervised Learning: Classification

- **Assumption 1:** The target variable (output) y is now binary
- **Assumption 2:** The input data x is separable
- **Definition:** We will denote $\mathcal{C}$ the set of possible classes that y can take



Supervised Learning: Classification

- **Assumption 1:** The target variable (output) y is now binary
- **Assumption 2:** The input data x is separable
- **Definition:** We will denote $\mathcal{C}$ the set of possible classes that y can take

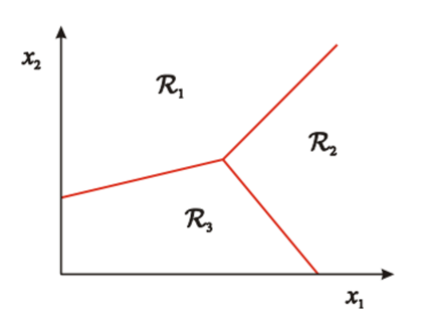


Goal:

To predict the correct class $y = c \in \mathcal{C}$ using x.
Alternative notation $y = \mathcal{C}_c$

Supervised Learning: Classification

- **Assumption 1:** The target variable (output) y is now binary
- **Assumption 2:** The input data x is separable
- **Definition:** We will denote $\mathcal{C}$ the set of possible classes that y can take



The boundaries of the **decision regions** are called **decision boundaries** or **surfaces**.

Goal:

To predict the correct class $y = c \in \mathcal{C}$ using x.
Alternative notation $y = \mathcal{C}_c$

The Learning Process

- We saw that the learning process if very similar to that one of regression

The Learning Process

- We saw that the learning process if very similar to that one of regression
- **but** we had to introduce a different loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{y}_i \neq y_i}, \quad \text{where} \quad \delta_{\hat{y}_i \neq y_i} = \begin{cases} 1, & \text{if} \quad \hat{y}_i \neq y_i \\ 0, & \text{otheriwse} \end{cases}$$

The Zero-one loss function

The Learning Process

- We saw that the learning process if very similar to that one of regression
- **but** we had to introduce a different loss function:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \delta_{\hat{y}_i \neq y_i}, \quad \text{where} \quad \delta_{\hat{y}_i \neq y_i} = \begin{cases} 1, & \text{if} \quad \hat{y}_i \neq y_i \\ 0, & \text{otheriwse} \end{cases}$$

The Zero-one loss function

- We derived an expression for the decision rule that minimizes the expected loss:

$$\min \sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$

The Bayes Classifier

- Using the 0-1 loss, we derived the Bayes Classifier.
- It classifies x to the most likely class using the conditional distribution

$$\begin{aligned}\hat{y} &= \arg\min_j 1 - p(\mathcal{C}_j|\mathbf{x}) \\ &= \arg\max_j p(\mathcal{C}_j|\mathbf{x})\end{aligned}$$

- It can be used as long as $p(\mathcal{C}_j|\mathbf{x})$ is known

The Bayes Classifier

- Using the 0-1 loss, we derived the Bayes Classifier.
- It classifies x to the most likely class using the conditional distribution

$$\begin{aligned}\hat{y} &= \arg\min_j 1 - p(\mathcal{C}_j|\mathbf{x}) \\ &= \arg\max_j p(\mathcal{C}_j|\mathbf{x})\end{aligned}$$

- It can be used as long as $p(\mathcal{C}_j|\mathbf{x})$ is known
- We introduced a method to estimate $p(\mathcal{C}_j|\mathbf{x})$:
  - Linear Discriminant Analysis - Generative approach

The Bayes Classifier

- Using the 0-1 loss, we derived the Bayes Classifier.
- It classifies x to the most likely class using the conditional distribution

$$\begin{aligned}\hat{y} &= \arg\min_j 1 - p(\mathcal{C}_j|\mathbf{x}) \\ &= \arg\max_j p(\mathcal{C}_j|\mathbf{x})\end{aligned}$$

- It can be used as long as $p(\mathcal{C}_j|\mathbf{x})$ is known
- We introduced a method to estimate $p(\mathcal{C}_j|\mathbf{x})$:
  - Linear Discriminant Analysis - Generative approach
- **Now:** A second method - Logistic regression (discriminative approach)

Logistic Regression

Probability Refresher: The Odds

- In statistics, **the odds for** or **odds of** some event reflect the likelihood that the event will take place, while **odds against** reflect the likelihood that it will not (Source: Wikipedia)

Probability Refresher: The Odds

- In statistics, **the odds for** or **odds of** some event reflect the likelihood that the event will take place, while **odds against** reflect the likelihood that it will not (Source: Wikipedia)
- Definition:

$$odds = \frac{\text{Probability of an event to occur}}{\text{Probability of not occuring}}$$

  often expressed as (Prob. of event occurence):(Prob. of non ocurrence)

Probability Refresher: The Odds

- In statistics, **the odds for** or **odds of** some event reflect the likelihood that the event will take place, while **odds against** reflect the likelihood that it will not (Source: Wikipedia)
- Definition:

$$odds = \frac{\text{Probability of an event to occur}}{\text{Probability of not occuring}}$$

  often expressed as (Prob. of event occurence):(Prob. of non ocurrence)
- Mathematically, it is a **Bernoulli trial** as it has two outcomes

Probability Refresher: The Odds

- In statistics, **the odds for** or **odds of** some event reflect the likelihood that the event will take place, while **odds against** reflect the likelihood that it will not (Source: Wikipedia)
- Definition:

$$odds = \frac{\text{Probability of an event to occur}}{\text{Probability of not occuring}}$$

  often expressed as (Prob. of event occurence):(Prob. of non ocurrence)
- Mathematically, it is a **Bernoulli trial** as it has two outcomes
- Example: Odds that a randomly chosen day of the week is a weekend

Probability Refresher: The Odds

- In statistics, **the odds for** or **odds of** some event reflect the likelihood that the event will take place, while **odds against** reflect the likelihood that it will not (Source: Wikipedia)
- Definition:

$$odds = \frac{\text{Probability of an event to occur}}{\text{Probability of not occuring}}$$

  often expressed as (Prob. of event occurence):(Prob. of non ocurrence)
- Mathematically, it is a **Bernoulli trial** as it has two outcomes
- Example: Odds that a randomly chosen day of the week is a weekend
- What about chances of choosing a day that is a weekend?

Odds & Probabilities: Relationship[1]

- Odds can be expressed as a ratio of two numbers:

$$1:1 \quad \text{or} \quad 100:100,$$

  which leads to a non-unique representation

[1] Source: https://en.wikipedia.org/wiki/Odds

Odds & Probabilities: Relationship[1]

- Odds can be expressed as a ratio of two numbers:

$$1 : 1 \quad \text{or} \quad 100 : 100,$$

  which leads to a non-unique representation
- or as a number, by dividing the terms in the ratio (unique representation)

[1]Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Relationship[1]

- Odds can be expressed as a ratio of two numbers:

$$1 : 1 \quad \text{or} \quad 100 : 100,$$

  which leads to a non-unique representation
- or as a number, by dividing the terms in the ratio (unique representation)
- Odds and probabilities are related by simple formulas
- Odds range from 0 to infinity, whereas probabilities go from 0 to 1

[1] Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Relationship[2]

- Given the odds as the ratio $W : L$ (Wins:Losses), the odds in favor (as a number) o_f and odds against (as a number) o_a can be computed by:

[2]Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Relationship[2]

- Given the odds as the ratio $W : L$ (Wins:Losses), the odds in favor (as a number) o_f and odds against (as a number) o_a can be computed by:

$$
\begin{aligned}
o_f &= W/L, \\
o_a &= L/W, \\
o_f \cdot o_a &= 1
\end{aligned}
$$

[2]Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Relationship[2]

- Given the odds as the ratio $W : L$ (Wins:Losses), the odds in favor (as a number) o_f and odds against (as a number) o_a can be computed by:

$$\begin{aligned} o_f &= W/L, \\ o_a &= L/W, \\ o_f \cdot o_a &= 1 \end{aligned}$$

- Analogously, given odds as a ratio, the probability of success p or failure q can be computed by:

[2]Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Relationship[2]

- Given the odds as the ratio $W : L$ (Wins:Losses), the odds in favor (as a number) o_f and odds against (as a number) o_a can be computed by:

$$o_f = W/L,$$
$$o_a = L/W,$$
$$o_f \cdot o_a = 1$$

- Analogously, given odds as a ratio, the probability of success p or failure q can be computed by:

$$p = W/(W + L),$$
$$q = L/(W + L),$$
$$p + q = 1$$

[2]Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Relationship[3]

- Given a probability p, the odds as a ratio (success to failure) is:

$$p : q$$

[3] Source: https://en.wikipedia.org/wiki/Odds

Odds & Probabilities: Relationship[3]

- Given a probability p, the odds as a ratio (success to failure) is:

$$p : q$$

- The odds as numbers can be computed by

$$o_f = p/q = p/(1-p) = (1-q)/q,$$
$$o_a = q/p = q/(1-q) = (1-p)/p$$

[3] Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Relationship[3]

- Given a probability p, the odds as a ratio (success to failure) is:

$$p : q$$

- The odds as numbers can be computed by

$$o_f = p/q = p/(1-p) = (1-q)/q,$$
$$o_a = q/p = q/(1-q) = (1-p)/p$$

- Given the odds as a number o_f, the ratio is expressed as $o_f : 1$ (conversely $1 : (1/o_f) = 1 : o_a$) from which p (conversely q) can be computed:

[3]Source: https://en.wikipedia.org/wiki/Odds

Odds & Probabilities: Relationship[3]

- Given a probability p, the odds as a ratio (success to failure) is:

$$p : q$$

- The odds as numbers can be computed by

$$o_f = p/q = p/(1-p) = (1-q)/q,$$
$$o_a = q/p = q/(1-q) = (1-p)/p$$

- Given the odds as a number o_f, the ratio is expressed as $o_f : 1$ (conversely $1 : (1/o_f) = 1 : o_a$) from which p (conversely q) can be computed:

$$p = o_f/(o_f + 1) = 1/(o_a + 1),$$
$$q = o_a/(o_a + 1) = 1/(o_f + 1)$$

[3]Source: `https://en.wikipedia.org/wiki/Odds`

Odds & Probabilities: Summary Table

odds (ratio)	o_f	o_a	p	q
1:1	1	1	50%	50%
0:1	0	∞	0%	100%
1:0	∞	0	100%	0%
2:1	2	0.5	67%	33%
1:2	0.5	2	33%	67%
4:1	4	0.25	80%	20%
1:4	0.25	4	20%	80%
9:1	9	$0.\overline{1}$	90%	10%
10:1	10	0.1	$90.\overline{90}\%$	$9.\overline{09}\%$
99:1	99	$0.\overline{01}$	99%	1%
100:1	100	0.01	$99.\overline{0099}\%$	$0.\overline{9900}\%$

Source: `https://en.wikipedia.org/wiki/Odds`

Linear Discriminant Analysis: Quick recap

- **Assumption:** Data within each class is normally distributed

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma})$$

Linear Discriminant Analysis: Quick recap

- **Assumption:** Data within each class is normally distributed

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma})$$

- For simplicity, let us assume we are in a two class setting, $K = 2$.
- LDA allowed us to compute the posteriors $p(\mathcal{C}_1|\mathbf{x})$, $p(\mathcal{C}_2|\mathbf{x})$

Linear Discriminant Analysis: Quick recap

- **Assumption:** Data within each class is normally distributed

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma})$$

- For simplicity, let us assume we are in a two class setting, $K = 2$.
- LDA allowed us to compute the posteriors $p(\mathcal{C}_1|\mathbf{x})$, $p(\mathcal{C}_2|\mathbf{x})$

- **Question:** Using what we saw about odds and probabilities, which are the odds for class 1?

Linear Discriminant Analysis: Quick recap

- **Assumption:** Data within each class is normally distributed

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma})$$

- For simplicity, let us assume we are in a two class setting, $K = 2$.
- LDA allowed us to compute the posteriors $p(\mathcal{C}_1|\mathbf{x})$, $p(\mathcal{C}_2|\mathbf{x})$

- **Question:** Using what we saw about odds and probabilities, which are the odds for class 1?

$$\frac{P(E)}{P(\overline{E})} = \frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}$$

Linear Discriminant Analysis: Quick recap

- **Assumption:** Data within each class is normally distributed

$$p(\mathbf{x}|\mathcal{C}_k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma})$$

- For simplicity, let us assume we are in a two class setting, $K = 2$.
- LDA allowed us to compute the posteriors $p(\mathcal{C}_1|\mathbf{x})$, $p(\mathcal{C}_2|\mathbf{x})$

- **Question:** Using what we saw about odds and probabilities, which are the odds for class 1?

$$\frac{P(E)}{P(\bar{E})} = \frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}$$

- As the log is monotonic we can estimate the log-odds:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right)$$

The Log-odds

Using the Bayes' theorem:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)/p(\mathbf{x})}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)/p(\mathbf{x})}\right)$$

The Log-odds

Using the Bayes' theorem:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)/p(\mathbf{x})}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)/p(\mathbf{x})}\right)$$

$$= \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$$

The Log-odds

Using the Bayes' theorem:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)/p(\mathbf{x})}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)/p(\mathbf{x})}\right)$$
$$= \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$$

Since the data within each class is normally distributed and recalling the notation from LDA where $\pi_k = p(\mathcal{C}_k)$, we get:

The Log-odds

Using the Bayes' theorem:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)/p(\mathbf{x})}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)/p(\mathbf{x})}\right)$$
$$= \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$$

Since the data within each class is normally distributed and recalling the notation from LDA where $\pi_k = p(\mathcal{C}_k)$, we get:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{\mathcal{N}(\mu_1, \mathbf{\Sigma})}{\mathcal{N}(\mu_2, \mathbf{\Sigma})}\right) + \log\left(\frac{\pi_1}{\pi_2}\right)$$

The Log-odds

Using the Bayes' theorem:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)/p(\mathbf{x})}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)/p(\mathbf{x})}\right)$$
$$= \log\left(\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)}\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$$

Since the data within each class is normally distributed and recalling the notation from LDA where $\pi_k = p(\mathcal{C}_k)$, we get:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{\mathcal{N}(\mu_1, \mathbf{\Sigma})}{\mathcal{N}(\mu_2, \mathbf{\Sigma})}\right) + \log\left(\frac{\pi_1}{\pi_2}\right)$$
$$= \log\left(\frac{C\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)\right)}{C\exp\left(-\frac{1}{2}(\mathbf{x}-\mu_2)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_2)\right)}\right) + \log\left(\frac{\pi_1}{\pi_2}\right)$$

C : Common constant value to both terms. Why?

The Log-odds: Derivation

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = -\frac{1}{2}\left(\mathbf{x}-\mu_1\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\mu_1\right) + \frac{1}{2}\left(\mathbf{x}-\mu_2\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\mu_2\right) + \log\left(\frac{\pi_1}{\pi_2}\right)$$

The Log-odds: Derivation

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = -\frac{1}{2}\left(\mathbf{x}-\mu_1\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\mu_1\right) + \frac{1}{2}\left(\mathbf{x}-\mu_2\right)^T \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\mu_2\right) + \log\left(\frac{\pi_1}{\pi_2}\right)$$

Exercise: Prove that

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}\left(\mu_1+\mu_2\right)^T \boldsymbol{\Sigma}^{-1}\left(\mu_1-\mu_2\right) + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\left(\mu_1-\mu_2\right)$$

The Log-odds: Derivation

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = -\frac{1}{2}\left(\mathbf{x}-\mu_1\right)^T \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mu_1\right) + \frac{1}{2}\left(\mathbf{x}-\mu_2\right)^T \mathbf{\Sigma}^{-1}\left(\mathbf{x}-\mu_2\right) + \log\left(\frac{\pi_1}{\pi_2}\right)$$

$$\begin{aligned}\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) &= -\frac{1}{2}\left(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\ldots\right) \\ &\quad + \frac{1}{2}\left(\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\ldots\right) + \log\left(\frac{\pi_1}{\pi_2}\right)\end{aligned}$$

Exercise: Prove that

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}\left(\mu_1+\mu_2\right)^T \mathbf{\Sigma}^{-1}\left(\mu_1-\mu_2\right) + \mathbf{x}^T\mathbf{\Sigma}^{-1}\left(\mu_1-\mu_2\right)$$

Motivation

Having a closer look at the expression we just obtained:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}(\mu_1 + \mu_2)^T \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2) + \mathbf{x}^T \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_2)$$

Motivation

Having a closer look at the expression we just obtained:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}\left(\mu_1 + \mu_2\right)^T \boldsymbol{\Sigma}^{-1}\left(\mu_1 - \mu_2\right) + \mathbf{x}^T \boldsymbol{\Sigma}^{-1}\left(\mu_1 - \mu_2\right)$$

Instead of estimating μ, $\boldsymbol{\Sigma}$ and π, we can directly estimate $\hat{w}_0$ and $\hat{\mathbf{w}}$.

Motivation

Having a closer look at the expression we just obtained:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) - \frac{1}{2}(\mu_1+\mu_2)^T\,\boldsymbol{\Sigma}^{-1}(\mu_1-\mu_2) + \mathbf{x}^T\boldsymbol{\Sigma}^{-1}(\mu_1-\mu_2)$$

Instead of estimating μ, $\boldsymbol{\Sigma}$ and π, we can directly estimate $\hat{w}_0$ and $\hat{\mathbf{w}}$. This is:

Logistic Regression

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Proof:

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Goal:

Estimate coefficients $\mathbf{w}$

Proof:

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Goal:

Estimate coefficients $\mathbf{w}$

Proof:

$$\log\frac{p}{1-p} = w_0 + \mathbf{w}^T\mathbf{x}$$

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Goal:

Estimate coefficients $\mathbf{w}$

Proof:

$$\log\frac{p}{1-p} = w_0 + \mathbf{w}^T\mathbf{x}$$

$$\frac{p}{1-p} = \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)$$

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Goal:

Estimate coefficients $\mathbf{w}$

Proof:

$$\begin{aligned}
\log\frac{p}{1-p} &= w_0 + \mathbf{w}^T\mathbf{x} \\
\frac{p}{1-p} &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right) \\
p &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)(1-p)
\end{aligned}$$

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Goal:

Estimate coefficients $\mathbf{w}$

Proof:

$$\begin{aligned}
\log\frac{p}{1-p} &= w_0 + \mathbf{w}^T\mathbf{x} \\
\frac{p}{1-p} &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right) \\
p &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)(1-p) \\
p &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right) - \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)p
\end{aligned}$$

Model Assumptions

Assumption:

$$\log\left(\frac{p(\mathcal{C}_j|\mathbf{x})}{p(\mathcal{C}_k|\mathbf{x})}\right) = w_0 + \mathbf{w}^T\mathbf{x}$$

where

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$

Goal:

Estimate coefficients $\mathbf{w}$

Proof:

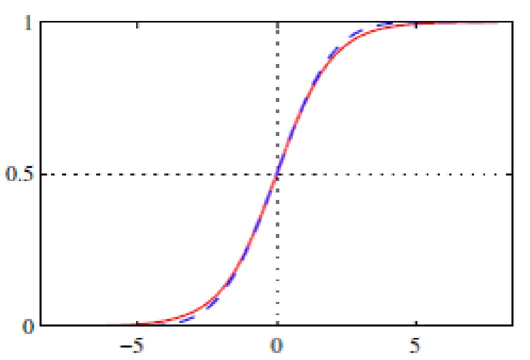
$$\begin{aligned}
\log\frac{p}{1-p} &= w_0 + \mathbf{w}^T\mathbf{x} \\
\frac{p}{1-p} &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right) \\
p &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)(1-p) \\
p &= \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right) - \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)p \\
p &= \frac{\exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)}{1 + \exp\left(w_0 + \mathbf{w}^T\mathbf{x}\right)}
\end{aligned}$$

with $p = p(\mathcal{C}_j|\mathbf{x})$

The Sigmoid function

Our model assumption about the posterior probabilities $p(\mathcal{C}_j|\mathbf{x})$ is represented by the sigmoid function:

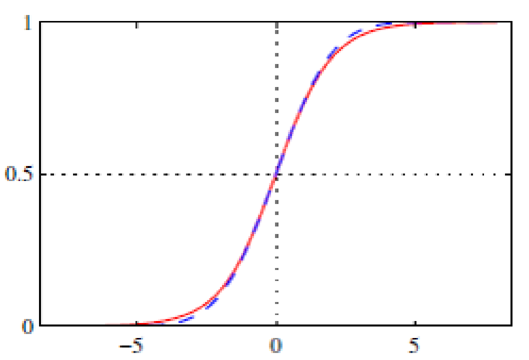
$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$



The Sigmoid function

Our model assumption about the posterior probabilities $p(\mathcal{C}_j|\mathbf{x})$ is represented by the sigmoid function:

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})} = \frac{1}{1 + \exp(-a)}$$

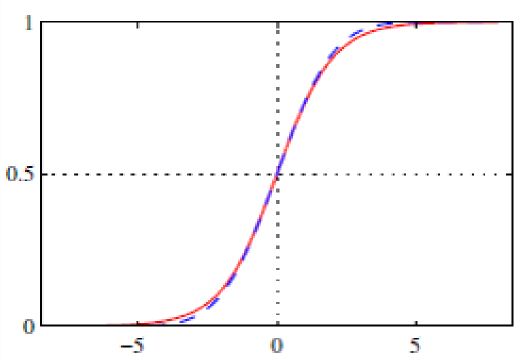


The Sigmoid function

Our model assumption about the posterior probabilities $p(\mathcal{C}_j|\mathbf{x})$ is represented by the sigmoid function:

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})}$$
$$= \frac{1}{1 + \exp(-a)}$$
$$= \sigma(a)$$

with $a = w_0 + \mathbf{w}^T\mathbf{x}$

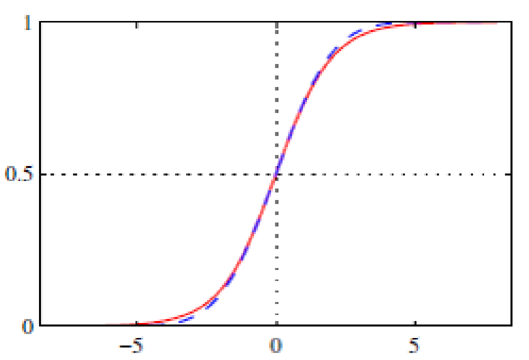


The Sigmoid function

Our model assumption about the posterior probabilities $p(\mathcal{C}_j|\mathbf{x})$ is represented by the sigmoid function:

$$p(\mathcal{C}_j|\mathbf{x}) = \frac{\exp(w_0 + \mathbf{w}^T\mathbf{x})}{1 + \exp(w_0 + \mathbf{w}^T\mathbf{x})} = \frac{1}{1 + \exp(-a)} = \sigma(a)$$

with $a = w_0 + \mathbf{w}^T\mathbf{x}$



Sigmoid means S-shaped. Also called squashing function because it maps the whole real axis into a finite interval

Properties of the Sigmoid Function

- Symmetry:

$$\sigma(-a) = 1 - \sigma(a)$$

Properties of the Sigmoid Function

- Symmetry:

$$\sigma(-a) = 1 - \sigma(a)$$

- Inverse or logit function:

$$a = \log\left(\frac{\sigma}{1-\sigma}\right)$$

Properties of the Sigmoid Function

- Symmetry:

$$\sigma(-a) = 1 - \sigma(a)$$

- Inverse or logit function:

$$a = \log\left(\frac{\sigma}{1-\sigma}\right)$$

Note that the inverse represents the log ratio of probabilities, which is nothing else than the log-odds:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})}\right)$$

Data Assumptions

- It is convenient to code the classes $\mathcal{C}_j$, via a $0/1$ response y:
  - $y_i = 1$ if $\mathcal{C}_1$
  - $y_i = 0$ if $\mathcal{C}_2$

Data Assumptions

- It is convenient to code the classes $\mathcal{C}_j$, via a $0/1$ response y:
  - $y_i = 1$ if $\mathcal{C}_1$
  - $y_i = 0$ if $\mathcal{C}_2$
- $y \in \{0, 1\}$ and $\mathbf{x} \in \mathbb{R}^D$

Data Assumptions

- It is convenient to code the classes $\mathcal{C}_j$, via a 0/1 response y:
    - $y_i = 1$ if $\mathcal{C}_1$
    - $y_i = 0$ if $\mathcal{C}_2$
- $y \in \{0, 1\}$ and $\mathbf{x} \in \mathbb{R}^D$
- The y_i 's are independent given the input features $\mathbf{x}_i$ and $\mathbf{w}$.

The Learning Process

- We will make use of the Maximum Likelihood Estimator (MLE) to fit our model.
- In other words, using a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, we use MLE to choose parameters that maximize the conditional likelihood:

The Learning Process

- We will make use of the Maximum Likelihood Estimator (MLE) to fit our model.
- In other words, using a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, we use MLE to choose parameters that maximize the conditional likelihood:

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{y}|\mathbf{X};\mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i;\mathbf{w})$$

The Learning Process

- We will make use of the Maximum Likelihood Estimator (MLE) to fit our model.
- In other words, using a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, we use MLE to choose parameters that maximize the conditional likelihood:

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i; \mathbf{w})$$

or rather the **log-likelihood** as it is easier to deal with:

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w})$$

The Learning Process

- We will make use of the Maximum Likelihood Estimator (MLE) to fit our model.
- In other words, using a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, we use MLE to choose parameters that maximize the conditional likelihood:

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{y}|\mathbf{X}; \mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i; \mathbf{w})$$

or rather the **log-likelihood** as it is easier to deal with:

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i; \mathbf{w})$$

- From what we saw in the first lecture about MLE, we just need to replace $p(y_i|\mathbf{x}_i; \mathbf{w})$ accordingly.

The Learning Process

- We will make use of the Maximum Likelihood Estimator (MLE) to fit our model.
- In other words, using a training dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, we use MLE to choose parameters that maximize the conditional likelihood:

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{y}|\mathbf{X};\mathbf{w}) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i;\mathbf{w})$$

  or rather the **log-likelihood** as it is easier to deal with:

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i;\mathbf{w})$$

- From what we saw in the first lecture about MLE, we just need to replace $p(y_i|\mathbf{x}_i;\mathbf{w})$ accordingly.
- **Question:** How?

The Learning Process (2)

1. We just saw that:
    - $y_i = 1 \Rightarrow \mathcal{C}_1$
    - $y_i = 0 \Rightarrow \mathcal{C}_2$

The Learning Process (2)

1. We just saw that:
    - $y_i = 1 \Rightarrow \mathcal{C}_1$
    - $y_i = 0 \Rightarrow \mathcal{C}_2$
2. We know that (for two classes):

$$p(\mathcal{C}_2|\mathbf{x}_i;\mathbf{w}) = 1 - p(\mathcal{C}_1|\mathbf{x}_i;\mathbf{w})$$

The Learning Process (2)

1. We just saw that:
   - $y_i = 1 \Rightarrow \mathcal{C}_1$
   - $y_i = 0 \Rightarrow \mathcal{C}_2$
2. We know that (for two classes):

$$p(\mathcal{C}_2|\mathbf{x}_i;\mathbf{w}) = 1 - p(\mathcal{C}_1|\mathbf{x}_i;\mathbf{w})$$

3. and we made some assumptions about the underlying model for $p(\mathcal{C}_j|\mathbf{x}_i;\mathbf{w})$.

Let's put all together to derive an expression for the log-likelihood in terms of $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{w}$.

The Learning Process (3)

- From point 1, we can rewrite the log-likelihood:

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} y_i \log p(\mathcal{C}_1 | \mathbf{x}_i; \mathbf{w}) + (1 - y_i) \log p(\mathcal{C}_2 | \mathbf{x}_i; \mathbf{w})$$

The Learning Process (3)

- From point 1, we can rewrite the log-likelihood:

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} y_i \log p(\mathcal{C}_1|\mathbf{x}_i;\mathbf{w}) + (1-y_i)\log p(\mathcal{C}_2|\mathbf{x}_i;\mathbf{w})$$

- From point 2, we can write everything in terms of $p(\mathcal{C}_1|\mathbf{x}_i;\mathbf{w})$:

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} y_i \log p(\mathcal{C}_1|\mathbf{x}_i;\mathbf{w}) + (1-y_i)\log\left(1 - p(\mathcal{C}_1|\mathbf{x}_i;\mathbf{w})\right)$$

- Finally, thanks to point 3, we make use of our model assumption:

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} y_i \log \sigma(w_0 + \mathbf{w}^T\mathbf{x}_i) + (1-y_i)\log\left(1 - \sigma(w_0 + \mathbf{w}^T\mathbf{x}_i)\right)$$

Cross-Entropy Loss Function

Instead of maximizing, we can take the negative of the previous expression to obtain a loss or error function[4]:

$$E(\mathbf{w}) = -\sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)\right) \tag{1}$$

This expression is denoted the **cross-entropy loss function**

[4]**Note:** We will assume the inputs $\mathbf{x}_i$ include a constant term 1 to be able to contract $\mathbf{w} = \{w_0, \mathbf{w}\}$

Cross-Entropy Loss Function

Instead of maximizing, we can take the negative of the previous expression to obtain a loss or error function[4]:

$$E(\mathbf{w}) = -\sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)\right) \tag{1}$$

This expression is denoted the **cross-entropy loss function**

To obtain the optimal $\mathbf{w}$, we need to derive the above expression and equal it to zero.

[4] **Note:** We will assume the inputs $\mathbf{x}_i$ include a constant term 1 to be able to contract $\mathbf{w} = \{w_0, \mathbf{w}\}$

Derivation

We will make use of the following property:

$$\frac{d\sigma}{da} = \sigma(1 - \sigma)$$

Deriving Eq. 1:

$$\frac{dE}{d\mathbf{w}} = -\frac{d}{d\mathbf{w}} \left(\sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)\right) \right)$$

Derivation

We will make use of the following property:

$$\frac{d\sigma}{da} = \sigma(1-\sigma)$$

Deriving Eq. 1:

$$\begin{aligned}
\frac{dE}{d\mathbf{w}} &= -\frac{d}{d\mathbf{w}}\left(\sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^T\mathbf{x}_i) + (1-y_i)\log\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\right) \\
&= -\sum_{i=1}^{N} y_i \frac{1}{\sigma(\mathbf{w}^T\mathbf{x}_i)} \cdot \sigma(\mathbf{w}^T\mathbf{x}_i)\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\mathbf{x}_i + \frac{d}{d\mathbf{w}}\left((1-y_i)\log\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\right)
\end{aligned}$$

Derivation

We will make use of the following property:

$$\frac{d\sigma}{da} = \sigma(1-\sigma)$$

Deriving Eq. 1:

$$\begin{aligned}
\frac{dE}{d\mathbf{w}} &= -\frac{d}{d\mathbf{w}}\left(\sum_{i=1}^{N} y_i \log \sigma(\mathbf{w}^T\mathbf{x}_i) + (1-y_i)\log\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\right) \\
&= -\sum_{i=1}^{N} y_i \frac{1}{\sigma(\mathbf{w}^T\mathbf{x}_i)} \cdot \sigma(\mathbf{w}^T\mathbf{x}_i)\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\mathbf{x}_i + \frac{d}{d\mathbf{w}}\left((1-y_i)\log\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\right) \\
&= -\sum_{i=1}^{N} y_i\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\mathbf{x}_i + (1-y_i)\frac{1}{1-\sigma(\mathbf{w}^T\mathbf{x}_i)} \cdot -\sigma(\mathbf{w}^T\mathbf{x}_i)\left(1-\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\mathbf{x}_i
\end{aligned}$$

Derivation (2)

$$\frac{dE}{d\mathbf{w}} = -\sum_{i=1}^{N} y_i \left(1 - \sigma(\mathbf{w}^T\mathbf{x}_i)\right) \mathbf{x}_i - (1 - y_i)\sigma(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i$$

Derivation (2)

$$\frac{dE}{d\mathbf{w}} = -\sum_{i=1}^{N} y_i \left(1 - \sigma(\mathbf{w}^T\mathbf{x}_i)\right)\mathbf{x}_i - (1 - y_i)\sigma(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i$$

$$= -\sum_{i=1}^{N} \left(y_i - y_i\sigma(\mathbf{w}^T\mathbf{x}_i) - \sigma(\mathbf{w}^T\mathbf{x}_i) + y_i\sigma(\mathbf{w}^T\mathbf{x}_i)\right)\mathbf{x}_i$$

Derivation (2)

$$
\begin{aligned}
\frac{dE}{d\mathbf{w}} &= -\sum_{i=1}^{N} y_i \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)\right) \mathbf{x}_i - (1 - y_i)\sigma(\mathbf{w}^T \mathbf{x}_i)\mathbf{x}_i \\
&= -\sum_{i=1}^{N} \left(y_i - y_i\sigma(\mathbf{w}^T \mathbf{x}_i) - \sigma(\mathbf{w}^T \mathbf{x}_i) + y_i\sigma(\mathbf{w}^T \mathbf{x}_i)\right) \mathbf{x}_i \\
&= -\sum_{i=1}^{N} \left(y_i - \cancel{y_i\sigma(\mathbf{w}^T \mathbf{x}_i)} - \sigma(\mathbf{w}^T \mathbf{x}_i) + \cancel{y_i\sigma(\mathbf{w}^T \mathbf{x}_i)}\right) \mathbf{x}_i
\end{aligned}
$$

Derivation (2)

$$
\begin{aligned}
\frac{dE}{d\mathbf{w}} &= -\sum_{i=1}^{N} y_i \left(1 - \sigma(\mathbf{w}^T\mathbf{x}_i)\right) \mathbf{x}_i - (1 - y_i)\sigma(\mathbf{w}^T\mathbf{x}_i)\mathbf{x}_i \\
&= -\sum_{i=1}^{N} \left(y_i - y_i\sigma(\mathbf{w}^T\mathbf{x}_i) - \sigma(\mathbf{w}^T\mathbf{x}_i) + y_i\sigma(\mathbf{w}^T\mathbf{x}_i)\right) \mathbf{x}_i \\
&= -\sum_{i=1}^{N} \left(y_i - \cancel{y_i\sigma(\mathbf{w}^T\mathbf{x}_i)} - \sigma(\mathbf{w}^T\mathbf{x}_i) + \cancel{y_i\sigma(\mathbf{w}^T\mathbf{x}_i)}\right) \mathbf{x}_i \\
&= \sum_{i=1}^{N} \left(\sigma(\mathbf{w}^T\mathbf{x}_i) - y_i\right) \mathbf{x}_i
\end{aligned}
$$

Derivation (2)

$$
\begin{aligned}
\frac{dE}{d\mathbf{w}} &= -\sum_{i=1}^{N} y_i \left(1 - \sigma(\mathbf{w}^T \mathbf{x}_i)\right) \mathbf{x}_i - (1 - y_i)\sigma(\mathbf{w}^T \mathbf{x}_i)\mathbf{x}_i \\
&= -\sum_{i=1}^{N} \left(y_i - y_i\sigma(\mathbf{w}^T \mathbf{x}_i) - \sigma(\mathbf{w}^T \mathbf{x}_i) + y_i\sigma(\mathbf{w}^T \mathbf{x}_i)\right) \mathbf{x}_i \\
&= -\sum_{i=1}^{N} \left(y_i - \cancel{y_i\sigma(\mathbf{w}^T \mathbf{x}_i)} - \sigma(\mathbf{w}^T \mathbf{x}_i) + \cancel{y_i\sigma(\mathbf{w}^T \mathbf{x}_i)}\right) \mathbf{x}_i \\
&= \sum_{i=1}^{N} \left(\sigma(\mathbf{w}^T \mathbf{x}_i) - y_i\right) \mathbf{x}_i
\end{aligned}
$$

Warning

After equating to zero, there is no way to obtain a closed-form solution for $\hat{\mathbf{w}}$

Predictions

- Formally, the parameters for logistic regression are estimated via:

$$\hat{w}_0, \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \; -\sum_{i=1}^{N} y_i \log \sigma(w_0 + \mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log \left(1 - \sigma(w_0 + \mathbf{w}^T \mathbf{x}_i)\right)$$

- We will not cover their estimation for now.

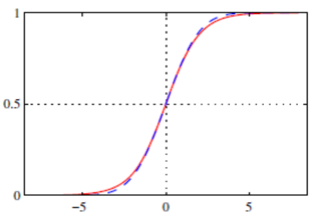
Predictions

- Formally, the parameters for logistic regression are estimated via:

$$\hat{w}_0, \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \; - \sum_{i=1}^{N} y_i \log \sigma(w_0 + \mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log \left(1 - \sigma(w_0 + \mathbf{w}^T \mathbf{x}_i)\right)$$

- We will not cover their estimation for now.
- Assuming we have $\hat{w}_0, \hat{\mathbf{w}}$, given a new sample $\mathbf{x}_{new}$, we will use what we know about the sigmoid function to assign y_{new}:

$$\begin{aligned} y_{new} &= p(\mathcal{C}_1 | \mathbf{x}_{new}; \hat{\mathbf{w}}) = \sigma(\hat{w}_0 + \hat{\mathbf{w}}^T \mathbf{x}_{new}) \\ &= \begin{cases} 1 & \text{if } \hat{w}_0 + \hat{\mathbf{w}}^T \mathbf{x}_{new} > 0 \\ 0 & \text{if } \hat{w}_0 + \hat{\mathbf{w}}^T \mathbf{x}_{new} \leq 0 \end{cases} \end{aligned}$$



Multiple Classes

- For K possible outcomes, we run $K-1$ independent binary logistic regression models

Multiple Classes

- For K possible outcomes, we run $K-1$ independent binary logistic regression models
- One outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_K|\mathbf{x})}\right) = w_{10} + \mathbf{w}_1^T\mathbf{x}$$

$$\dots$$

$$\log\left(\frac{p(\mathcal{C}_{K-1}|\mathbf{x})}{p(\mathcal{C}_K|\mathbf{x})}\right) = w_{(K-1)0} + \mathbf{w}_{K-1}^T\mathbf{x}$$

Multiple Classes

- For K possible outcomes, we run $K-1$ independent binary logistic regression models
- One outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_K|\mathbf{x})}\right) = w_{10} + \mathbf{w}_1^T\mathbf{x}$$

$$\cdots$$

$$\log\left(\frac{p(\mathcal{C}_{K-1}|\mathbf{x})}{p(\mathcal{C}_K|\mathbf{x})}\right) = w_{(K-1)0} + \mathbf{w}_{K-1}^T\mathbf{x}$$

where

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{\exp(w_{k0} + \mathbf{w}_k^T\mathbf{x})}{1 + \sum_{l=1}^{K-1}\exp(w_{l0} + \mathbf{w}_l^T\mathbf{x})}, \quad p(\mathcal{C}_K|\mathbf{x}) = \frac{1}{1 + \sum_{l=1}^{K-1}\exp(w_{l0} + \mathbf{w}_l^T\mathbf{x})}$$

Multiple Classes

- For K possible outcomes, we run $K-1$ independent binary logistic regression models
- One outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome:

$$\log\left(\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_K|\mathbf{x})}\right) = w_{10} + \mathbf{w}_1^T\mathbf{x}$$

$$\cdots$$

$$\log\left(\frac{p(\mathcal{C}_{K-1}|\mathbf{x})}{p(\mathcal{C}_K|\mathbf{x})}\right) = w_{(K-1)0} + \mathbf{w}_{K-1}^T\mathbf{x}$$

where

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{\exp(w_{k0} + \mathbf{w}_k^T\mathbf{x})}{1 + \sum_{l=1}^{K-1}\exp(w_{l0} + \mathbf{w}_l^T\mathbf{x})}, \quad p(\mathcal{C}_K|\mathbf{x}) = \frac{1}{1 + \sum_{l=1}^{K-1}\exp(w_{l0} + \mathbf{w}_l^T\mathbf{x})}$$

- **Exercise:** Verify that this is correct and that $\sum_k p(\mathcal{C}_k|\mathbf{x})$ sums to 1.

Recap

Recap

What we have seen so far...

- We introduced more formally the classification task
- We presented linear discriminant analysis (LDA)
- and we covered logistic regression
- Differently from LDA, it is a discriminative approach
- We saw that by relying in the log-odds logistic regression avoids to estimate the parameters of the generative model (as LDA)
- We have introduced the sigmoid function

What we have NOT seen...

- How to estimate the model parameters $\hat{\mathbf{w}}$ of the logistic regression model

References

Further Reading and Useful Material

Source	Notes
Pattern Recognition and Machine Learning	Ch. 4
The Elements of Statistical Learning	Ch. 2 and 4
Machine Learning - Tom Mitchel	Chapter 3 (link)