Logistic Regression

MGTF 495

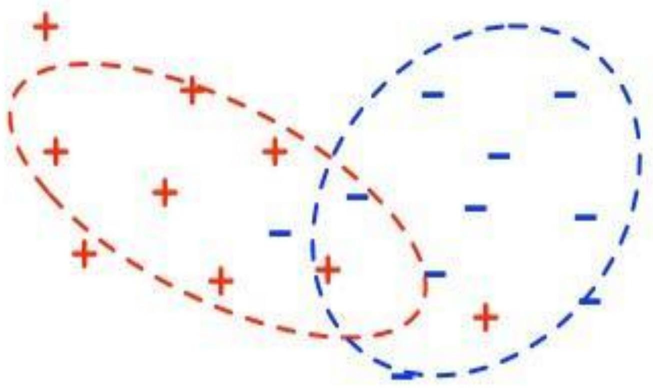
Class Outline

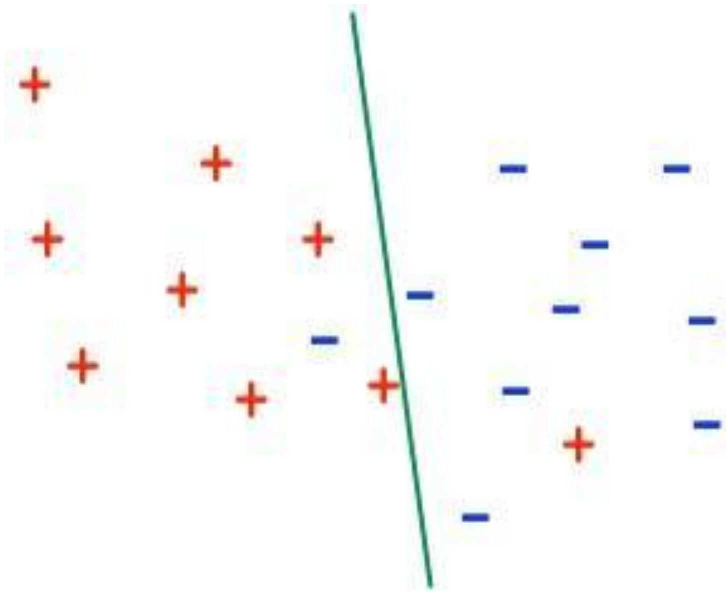
- Generative vs Discriminative Models
- Discriminative Models
    - Logistic Regression
    - SVM
    - Perceptron
- Kernels

Classification with parametrized models

Classifiers with a fixed no. of parameters can represent a limited set of functions. Learning a model is about picking a good approximation.

Typically the x 's are points in p-dimensional Euclidean space, R^p





Two ways to classify:

- **Generative**: model the individual classes.
- **Discriminative**: model the decision boundary between the classes.

Generative models: pros and cons

Advantages:

- Multiclass is a breeze
- For many common models: converges fast
- Returns not just a classification but also a confidence $\Pr(y|x)$

Generative models: pros and cons

Advantages:

- Multiclass is a breeze
- For many common models: converges fast
- Returns not just a classification but also a confidence $\Pr(y|x)$

Disadvantages:

Formula for $\Pr(y|x)$ assumes the class specific density models are perfect but this is never true

Generative models: pros and cons

Advantages:

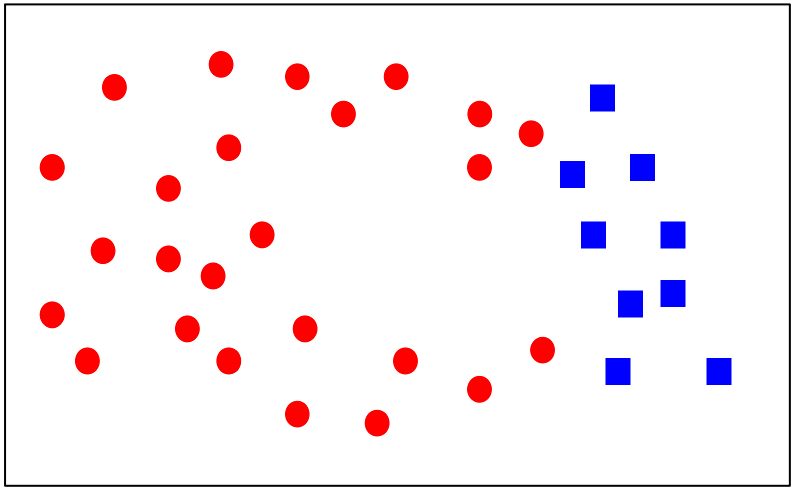
- Multiclass is a breeze
- For many common models: converges fast
- Returns not just a classification but also a confidence $\Pr(y|x)$

Disadvantages:

Formula for $\Pr(y|x)$ assumes the class specific density models are perfect but this is never true

If we only care about classification, shouldn't we focus on the decision boundary rather than trying to model other aspects of the distribution of x ?

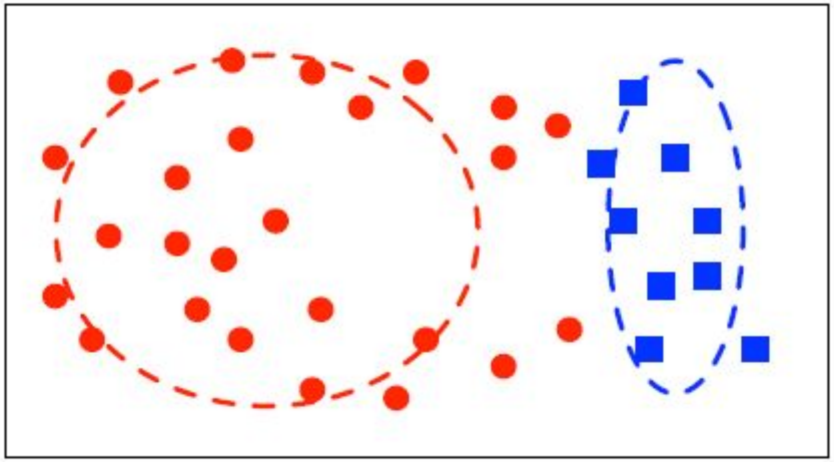
Generative versus discriminative



The generative way:

- Fit: π_0, π_1, P_0, P_1
- This determines a full joint distribution $\Pr(x, y)$
- Use Bayes' rule to obtain $\Pr(y|x)$

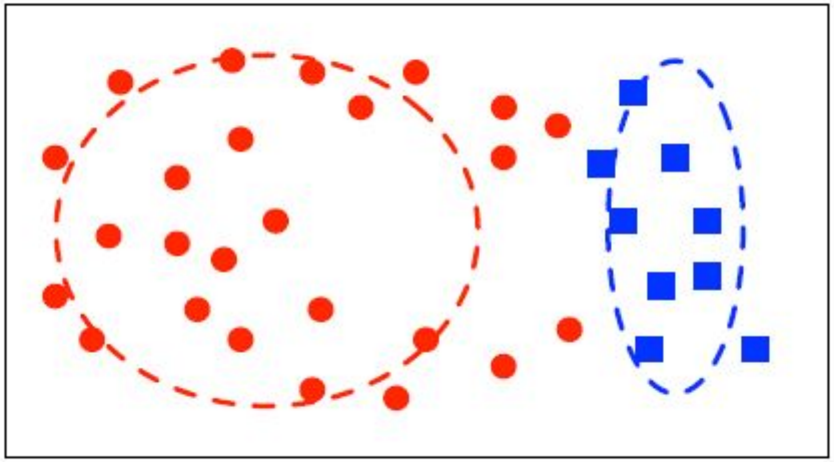
Generative versus discriminative



The generative way:

- Fit: π_0, π_1, P_0, P_1
- This determines a full joint distribution $\Pr(x, y)$
- Use Bayes' rule to obtain $\Pr(y|x)$

Generative versus discriminative



The generative way:

- Fit: π_0, π_1, P_0, P_1
- This determines a full joint distribution Pr(x , y)
- Use Bayes' rule to obtain Pr ($y|x$)

The generative way: model Pr($y|x$) directly.
In our earlier terminology: forget about the μ (Prob. Distribution), just learn the η (likelihood)

Class Outline

- Generative vs Discriminative Models
- Discriminative Models
  - Logistic Regression
  - SVM
  - Perceptron
- Kernels

The logistic regression model

What model to use for $\Pr(y \mid x)$?

The logistic regression model

What model to use for Pr(*y* |*x*)?

- Say $\mathcal{Y} = \{-1, 1\}$. Recall: for Gaussians with common covariance,

$$\ln \frac{\Pr(y = 1 \mid x)}{\Pr(y = -1 \mid x)} = \underbrace{w \cdot x + \theta}_{\text{linear}}$$

The logistic regression model

What model to use for Pr(y | x)?

- Say $\mathcal{Y} = \{-1, 1\}$. Recall: for Gaussians with common covariance,

$$\ln \frac{\Pr(y = 1 \mid x)}{\Pr(y = -1 \mid x)} = \underbrace{w \cdot x + \theta}_{\text{linear}}$$

- Can drop θ by adding an extra feature to x.

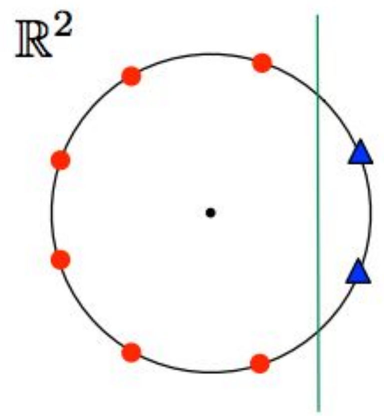
Homogeneous linear separators

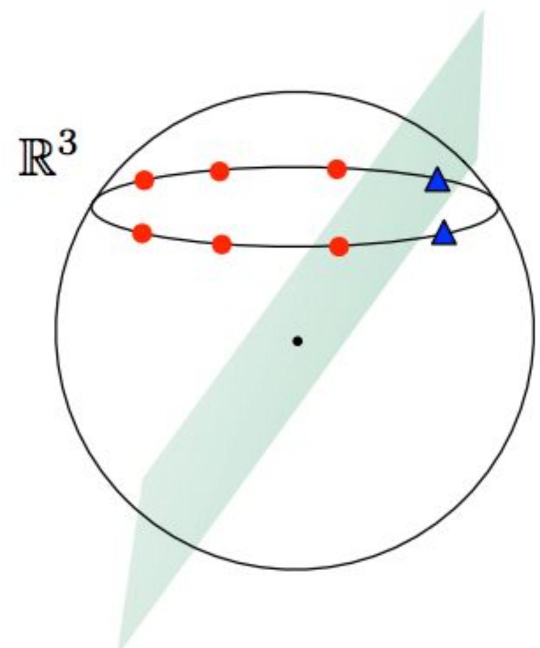
Hyperplanes that pass through the origin have no offset, $b = 0$.

Reduce to this case by adding an extra feature to x :

$$\widetilde{x} = (x, 1) \in \mathbb{R}^{p+1}$$

Then $\{x : w \cdot x = b\} \equiv \{x : \widetilde{w} \cdot \widetilde{x} = 0\}$ where $\widetilde{w} = (w, -b)$.





The logistic regression model

What model to use for $\Pr(y \mid x)$?

- Say $\mathcal{Y} = \{-1, 1\}$. Recall: for Gaussians with common covariance,

$$\ln \frac{\Pr(y = 1 \mid x)}{\Pr(y = -1 \mid x)} = \underbrace{w \cdot x + \theta}_{\text{linear}}$$

- Can drop θ by adding an extra feature to x.
- Then $\Pr(y = 1 \mid x) = \Pr(y = -1 \mid x)\, e^{w \cdot x}$, where
  $\Pr(y = 1 \mid x) = 1 - \Pr(y = -1 \mid x)$

The logistic regression model

What model to use for Pr(*y* |*x*)?

- Say $\mathcal{Y} = \{-1, 1\}$. Recall: for Gaussians with common covariance,

$$\ln \frac{\Pr(y = 1 \mid x)}{\Pr(y = -1 \mid x)} = \underbrace{w \cdot x + \theta}_{\text{linear}}$$

- Can drop θ by adding an extra feature to x.
- Then $\Pr(y = 1 \mid x) = \Pr(y = -1 \mid x)\, e^{w \cdot x}$, where

$\Pr(y = 1 \mid x) = 1 - \Pr(y = -1 \mid x)$

$$\Pr(y = -1 \mid x) = \frac{1}{1 + e^{w \cdot x}}$$

$$\Pr(y = 1 \mid x) = 1 - \frac{1}{1 + e^{w \cdot x}} = \frac{e^{w \cdot x}}{1 + e^{w \cdot x}} = \frac{1}{1 + e^{-w \cdot x}}$$

The logistic regression model

What model to use for Pr(y |x)?

- Say $\mathcal{Y} = \{-1, 1\}$. Recall: for Gaussians with common covariance,

$$\ln \frac{\Pr(y = 1 \mid x)}{\Pr(y = -1 \mid x)} = \underbrace{w \cdot x + \theta}_{\text{linear}}$$

- Can drop θ by adding an extra feature to x.
- Then $\Pr(y = 1 \mid x) = \Pr(y = -1 \mid x)\, e^{w \cdot x}$, where

$\Pr(y = 1 \mid x) = 1 - \Pr(y = -1 \mid x)$

$$\Pr(y = -1 \mid x) = \frac{1}{1 + e^{w \cdot x}}$$

$$\Pr(y = 1 \mid x) = 1 - \frac{1}{1 + e^{w \cdot x}} = \frac{e^{w \cdot x}}{1 + e^{w \cdot x}} = \frac{1}{1 + e^{-w \cdot x}}$$

- More concisely,

$$\Pr(y \mid x) = \frac{1}{1 + e^{-y(w \cdot x)}}$$

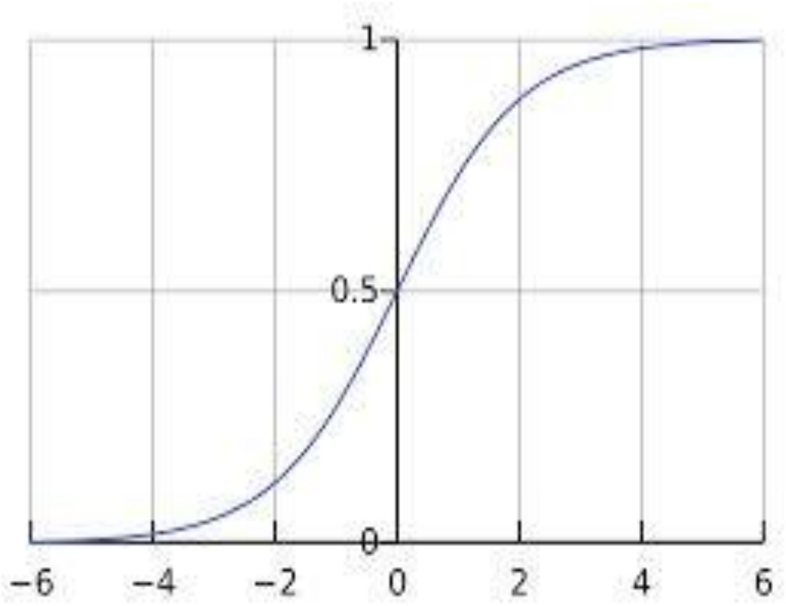
This is the **logistic regression model**, parametrized by w.

The squashing function

Take $X = \mathbf{R}^p$ and $Y = \{-1, 1\}$. The model specified by $w \in \mathbf{R}^p$ is

$$\Pr_w (y \mid x) = \frac{1}{1 + e^{-y(w \cdot x)}} = g(y(w \cdot x)),$$

where $g(z) = 1/(1 + e^{-z})$ is the *squashing function*.

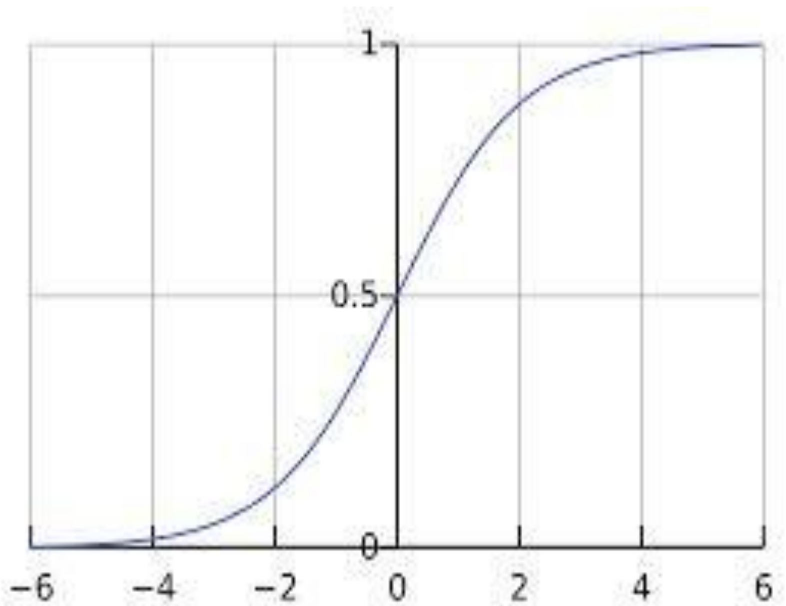


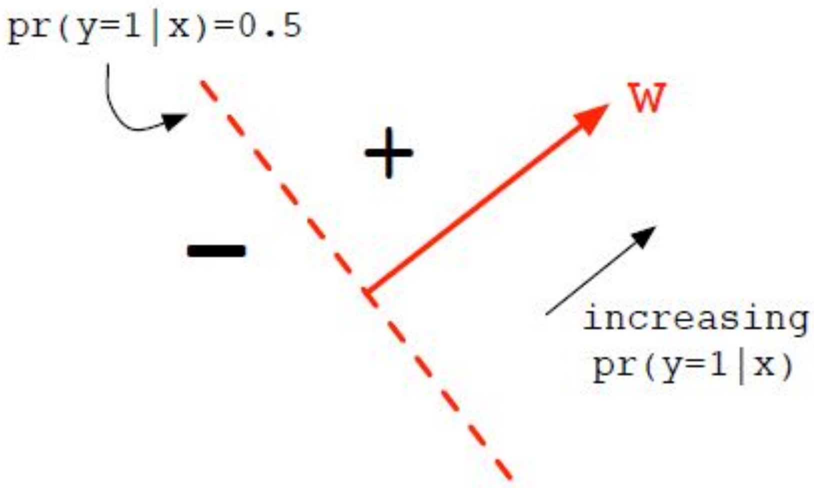
The squashing function

Take $X = \mathbf{R}^p$ and $Y = \{-1, 1\}$. The model specified by $w \in \mathbf{R}^p$ is

$$\Pr_w(y \mid x) = \frac{1}{1 + e^{-y(w \cdot x)}} = g(y(w \cdot x)),$$

where $g(z) = 1/(1 + e^{-z})$ is the *squashing function*.





Fitting *w*

The maximum-likelihood principle: given a data set

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^p \times \{-1, 1\},$$

pick the $w \in \mathbb{R}^p$ that maximizes

$$\prod_{i=1}^{n} \Pr_w(y^{(i)} \mid x^{(i)}).$$

Fitting w

The maximum-likelihood principle: given a data set

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^p \times \{-1, 1\},$$

pick the $w \in \mathbb{R}^p$ that maximizes

$$\prod_{i=1}^{n} \Pr_w(y^{(i)} \mid x^{(i)}).$$

Easier to work with sums, so take negative log to get **loss function**

$$L(w) = -\sum_{i=1}^{n} \ln \Pr_w(y^{(i)} \mid x^{(i)})$$

$$= -\sum_{i=1}^{n} \ln\left(\frac{1}{1 + e^{-y^{(i)}(w \cdot x^{(i)})}}\right) = \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$$

Our goal is to minimize $L(w)$.

Fitting w

The maximum-likelihood principle: given a data set

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^p \times \{-1, 1\},$$

pick the $w \in \mathbb{R}^p$ that maximizes

$$\prod_{i=1}^{n} \Pr_w(y^{(i)} \mid x^{(i)}).$$

Easier to work with sums, so take negative log to get **loss function**

$$L(w) \;=\; -\sum_{i=1}^{n} \ln \Pr_w(y^{(i)} \mid x^{(i)})$$

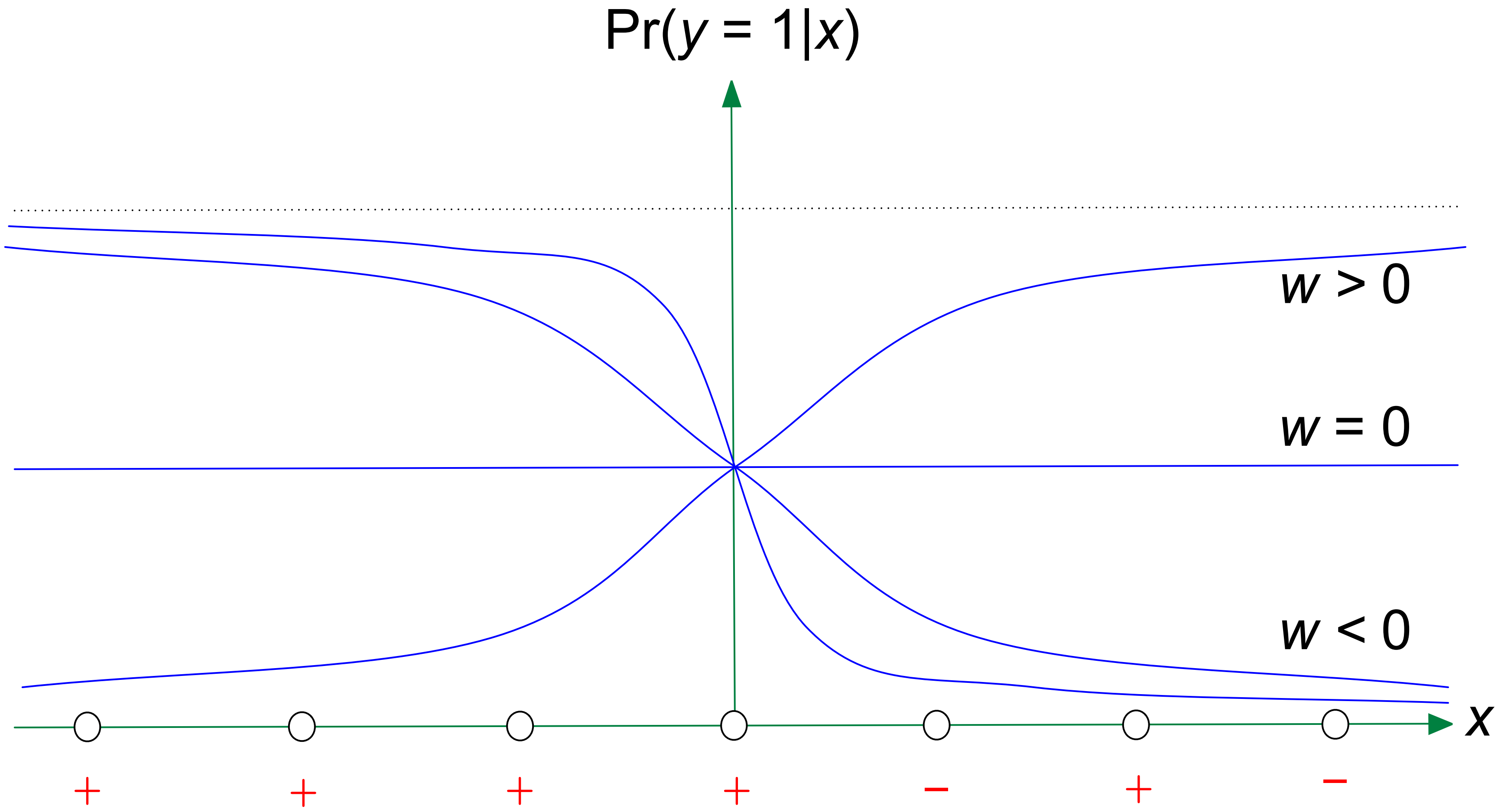
$$= \; -\sum_{i=1}^{n} \ln\left(\frac{1}{1 + e^{-y^{(i)}(w \cdot x^{(i)})}}\right) \;=\; \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}(w \cdot x^{(i)})}\right)$$

Our goal is to minimize $L(w)$.

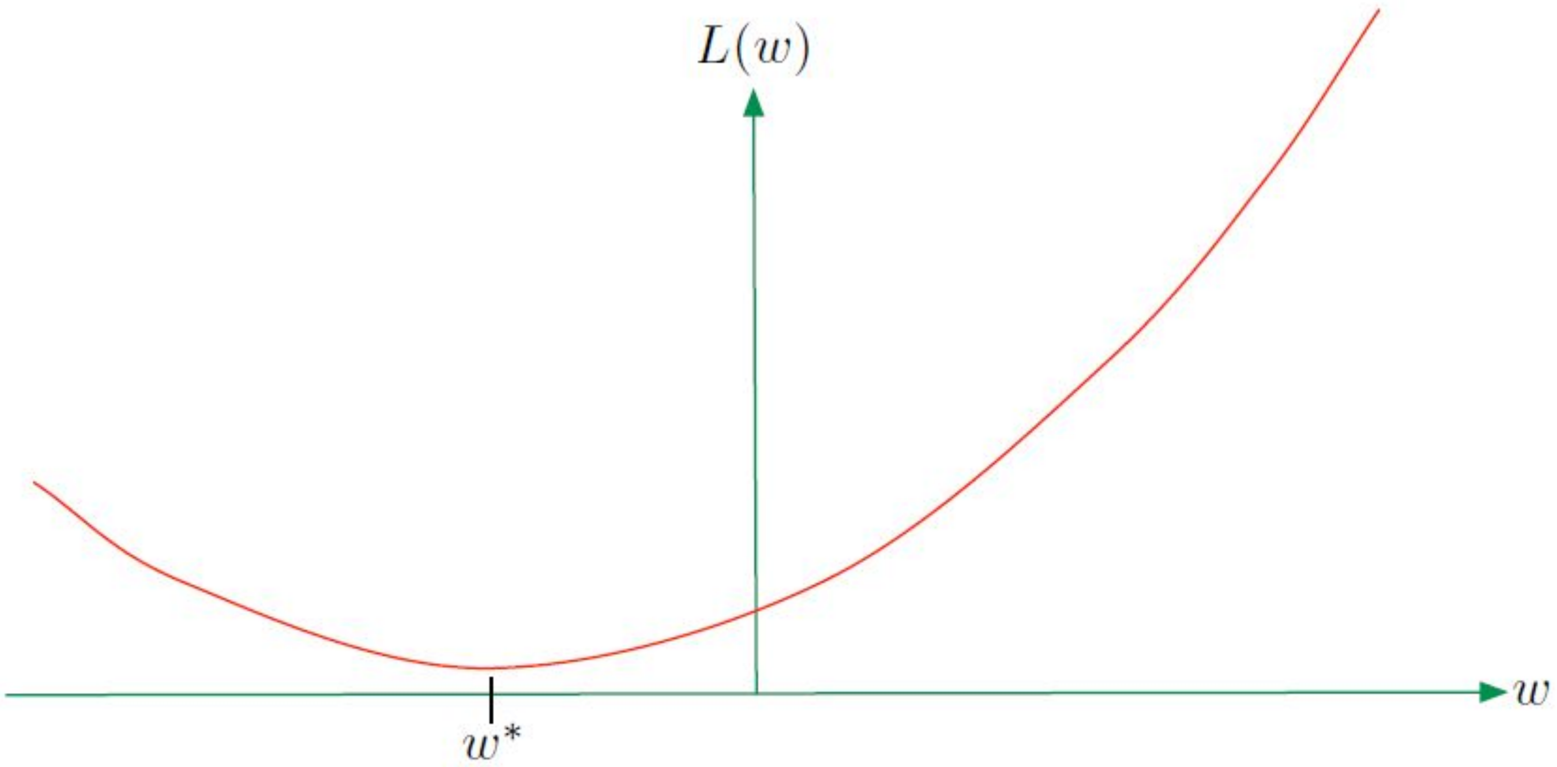
The good news: $L(w)$ is **convex** in w.

One dimensional example

$$\Pr_w(y \mid x) = \frac{1}{1 + e^{-ywx}}, \quad w \in \mathbf{R}$$



Example, cont'd



How to find the minimum of this convex function? A variety of options:

- Gradient descent
- Newton-Raphson

and many others.

Gradient descent procedure for LR

Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^p \times \{-1, 1\}$, find

$$\arg\min_{w \in \mathbb{R}^p} L(w) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$$

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots,$ until convergence:

$$w_{t+1} \;=\; w_t \;+\; \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \underbrace{\mathrm{Pr}_{w_t}(-y^{(i)} | x^{(i)})}_{\text{doubt}_t(x^{(i)}, y^{(i)})},$$

where η_t is a step size chosen by line search to minimize $L(w_{t+1})$.

Newton-Raphson procedure for LR

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots,$ until convergence:

$$w_{t+1} \;=\; w_t \;+\; \eta_t \,(X^T D_t X)^{-1} \sum_{i=1}^{n} y^{(i)} x^{(i)} \mathrm{Pr}_{w_t}(-y^{(i)}|x^{(i)}),$$

  where

  - X is the $n \times p$ data matrix with one point per row
  - D_t is an $n \times n$ diagonal matrix with (i, i) entry

$$D_{t,ii} \;=\; \mathrm{Pr}_{w_t}(1|x^{(i)})\,\mathrm{Pr}_{w_t}(-1|x^{(i)})$$

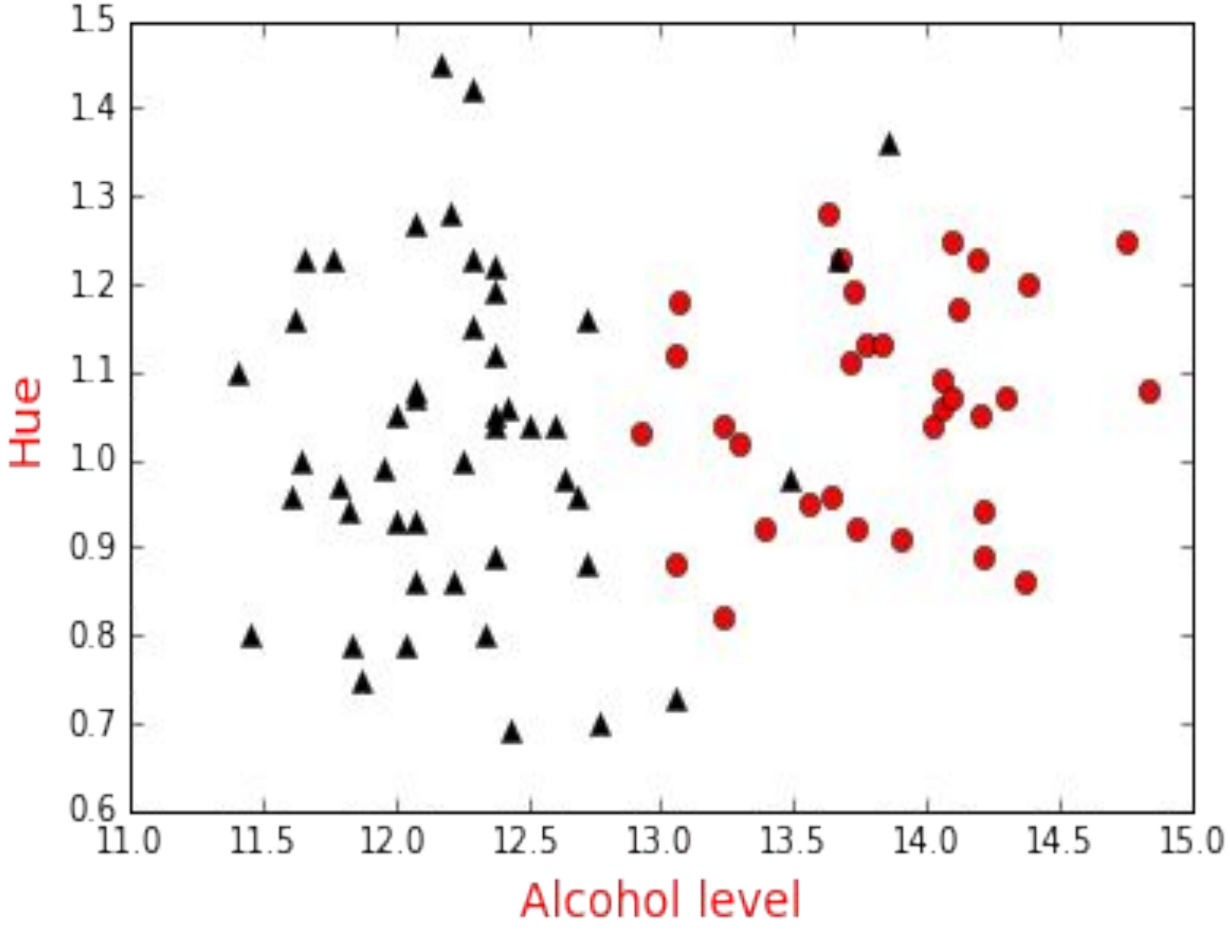
  - η_t is a step size that is either fixed to 1 ("iterative reweighted least squares") or chosen by line search to minimize $L(w_{t+1})$.

Example: "wine" data set

Recall: data from three wineries from the same region of Italy.

- 13 attributes: hue, color intensity, flavanoids, ash content, ...
- 178 instances in all: split into 118 train, 60 test

Pick two classes and just two attributes (hue, alcohol content).

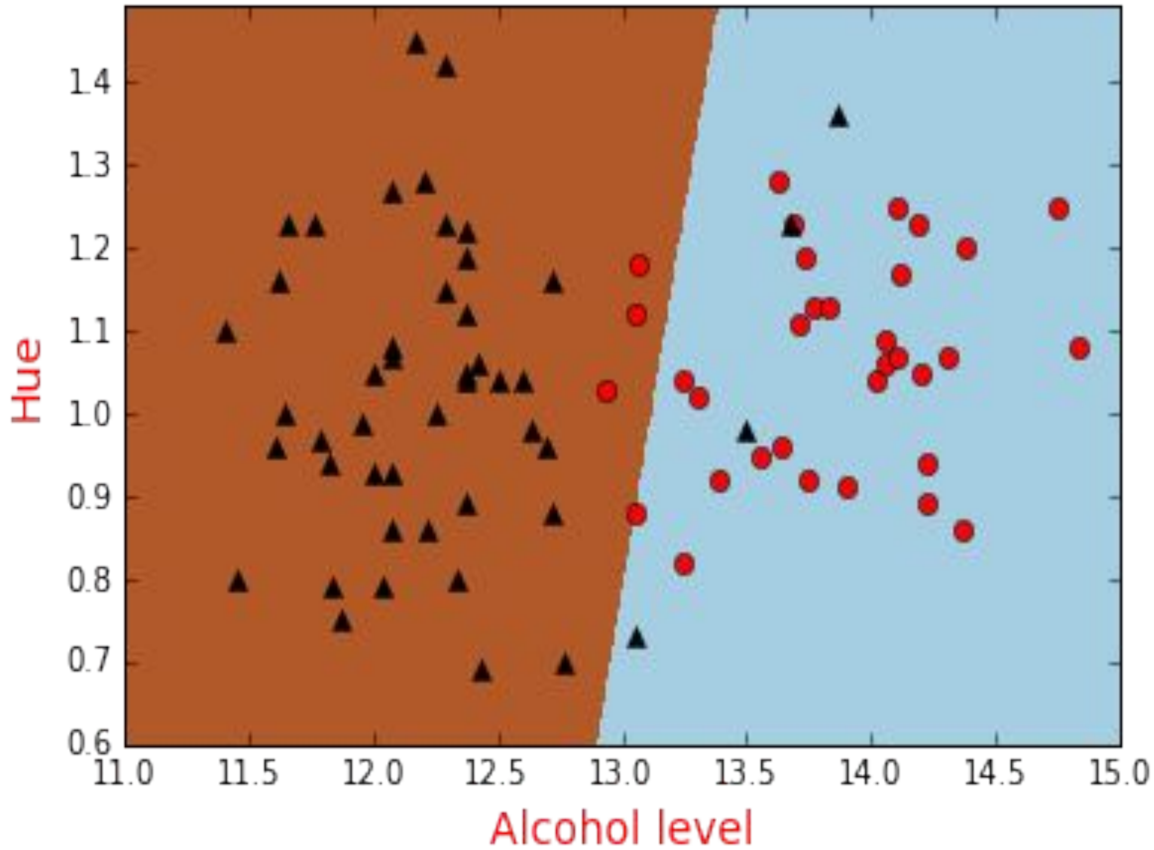


Example: "wine" data set

Recall: data from three wineries from the same region of Italy.

- 13 attributes: hue, color intensity, flavanoids, ash content, ...
- 178 instances in all: split into 118 train, 60 test

Pick two classes and just two attributes (hue, alcohol content).



Test error using logistic regression: 10%.