Lecture 6 – Logistic Regression

ECE597ML-697ML

Mario Parente

Logistic Regression

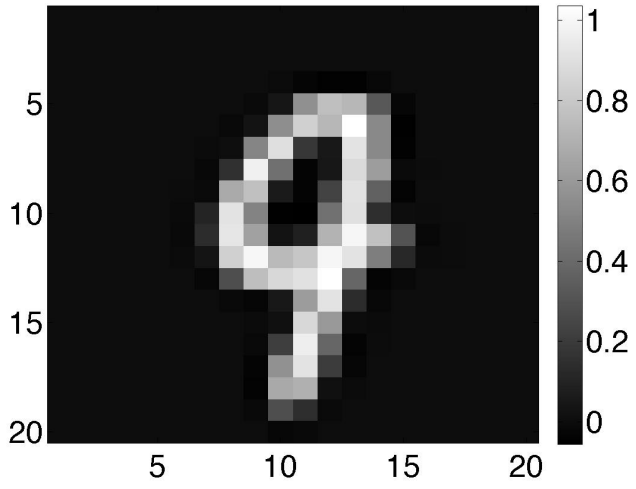
- Classification
- Model
- Cost function
- Gradient descent
- Linear classifiers and decision boundaries

Classification

- Input: $\mathbf{x} \in \mathbb{R}^n$
- Output: $y \in \{0, 1\}$

Example: Hand-Written Digits

Input: 20×20 grayscale image



$$\begin{bmatrix} x_1 & x_{21} & \dots & x_{381} \\ x_2 & x_{22} & \dots & x_{382} \\ & & \vdots & \\ x_{20} & x_{40} & \dots & x_{400} \end{bmatrix}$$

Unroll image into a feature vector $\mathbf{x} \in \mathbb{R}^{400}$

$$\mathbf{x} = (x_1, \dots, x_{400})^T$$

Output:

$$y = \begin{cases} 0 & \text{digit is "four"} \\ 1 & \text{digit is "nine"} \end{cases}$$

Example: Spam Classification

feature vector is composed by the frequency of occurrence of words

	class_label	message
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...
11	spam	SIX chances to win CASH! From 100 to 20,000 po...
...	...	...
5537	spam	Want explicit SEX in 30 secs? Ring 02073162414...
5540	spam	ASKED 3MOBILE IF 0870 CHATLINES INCLU IN FREE ...
5547	spam	Had your contract mobile 11 Mnths? Latest Moto...
5566	spam	REMINDER FROM O2: To get 2.50 pounds free call...
5567	spam	This is the 2nd time we have tried 2 contact u...

ham word cloud

spam word cloud

The Learning Problem

- Input: $\mathbf{x} \in \mathbb{R}^n$
- Output: $y \in \{0, 1\}$
- Model (hypothesis class): ?
- Cost function: ?

Classification as regression?

We could approach the classification problem ignoring the fact that y is discrete-valued, and use our old linear regression algorithm to try to predict y given $\mathbf{x}$. However, it is easy to construct examples where this method performs very poorly. Intuitively, it also doesn't make sense for $h_{\boldsymbol{\theta}}(\mathbf{x})$ to take values larger than 1 or smaller than 0 when we know that $y \in 0, 1$.

The Model

Exercise: fix the linear regression model

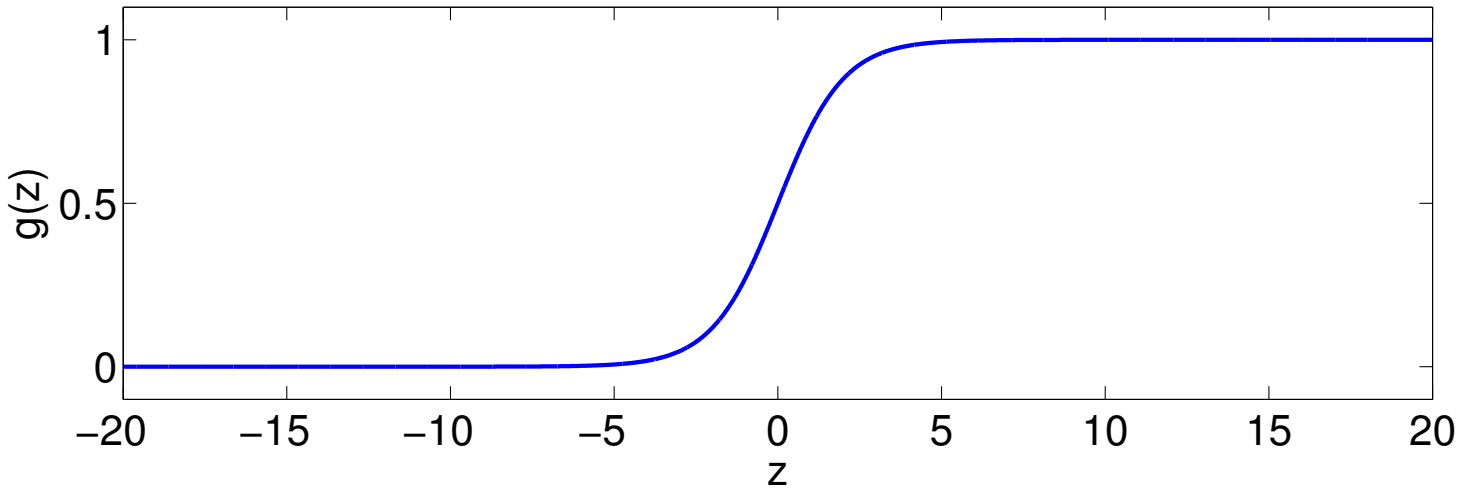
$$h_{\boldsymbol{\theta}}(\mathbf{x}) = g(\boldsymbol{\theta}^T \mathbf{x}), \qquad g : \mathbb{R} \to [0, 1].$$

where $\boldsymbol{\theta}^T \mathbf{x} = \theta_0 + \sum_{j=1}^{n} \theta_j x_j$.

What should g look like?

Logistic Function

$$g(z) = \frac{1}{1 + e^{-z}}$$



- This is called the *logistic* or *sigmoid* function

$$g(z) = \text{logistic}(z) = \text{sigmoid}(z)$$

The Model

Put it together

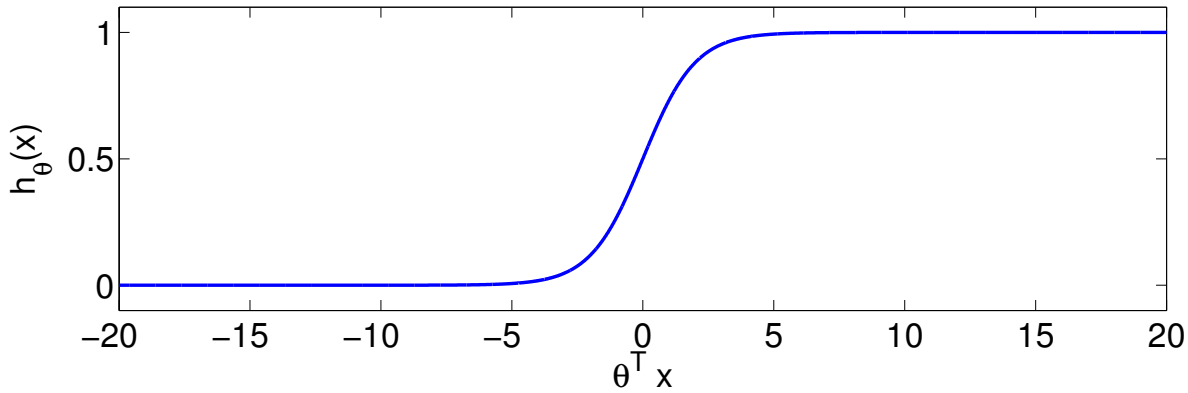
$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \text{logistic}(\boldsymbol{\theta}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}}$$

Nuance:

- Output is in $[0, 1]$, not $\{0, 1\}$.
- Interpret as probability

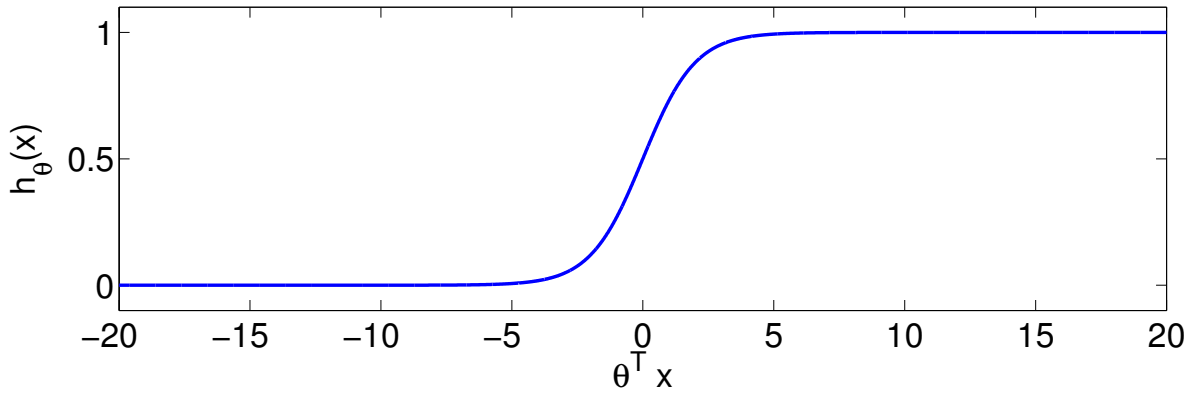
Hypothesis vs. Prediction Rule

Hypothesis (for learning, or when probability is useful)

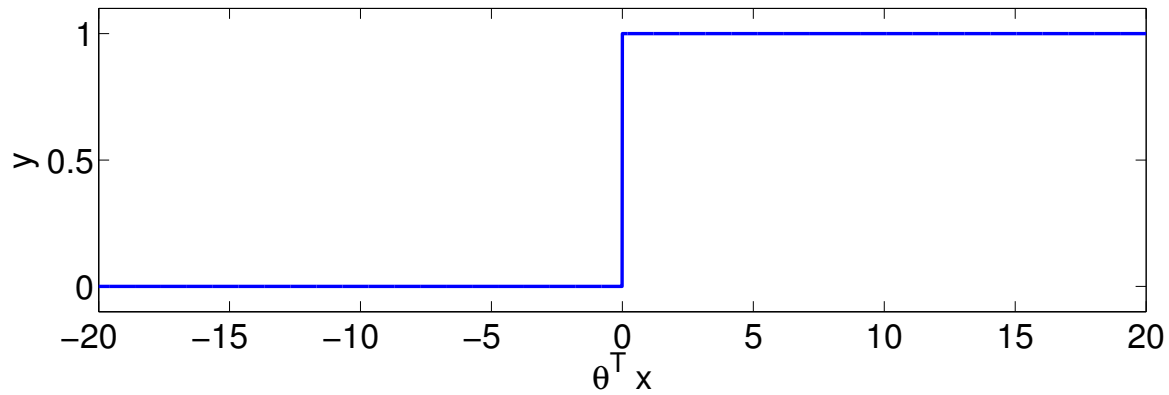


Hypothesis vs. Prediction Rule

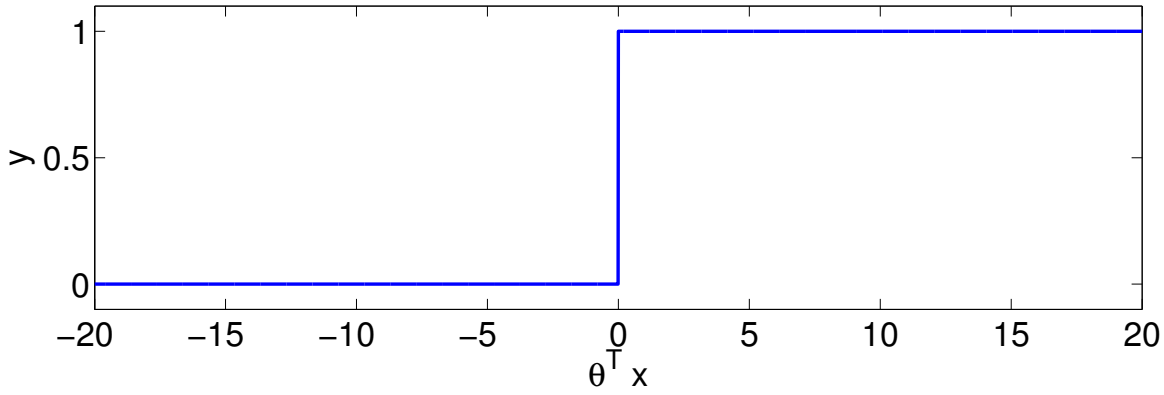
Hypothesis (for learning, or when probability is useful)



Prediction rule (when you need to commit!)



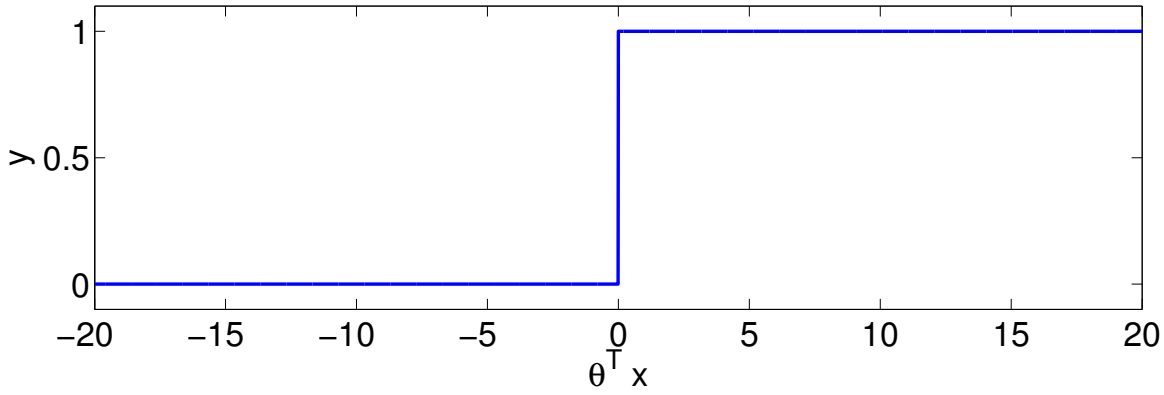
Prediction Rule



Rule

$$y = \begin{cases} 0 & \text{if } h_{\boldsymbol{\theta}}(\mathbf{x}) < 1/2 \\ 1 & \text{if } h_{\boldsymbol{\theta}}(\mathbf{x}) \geq 1/2 \end{cases}$$

Prediction Rule



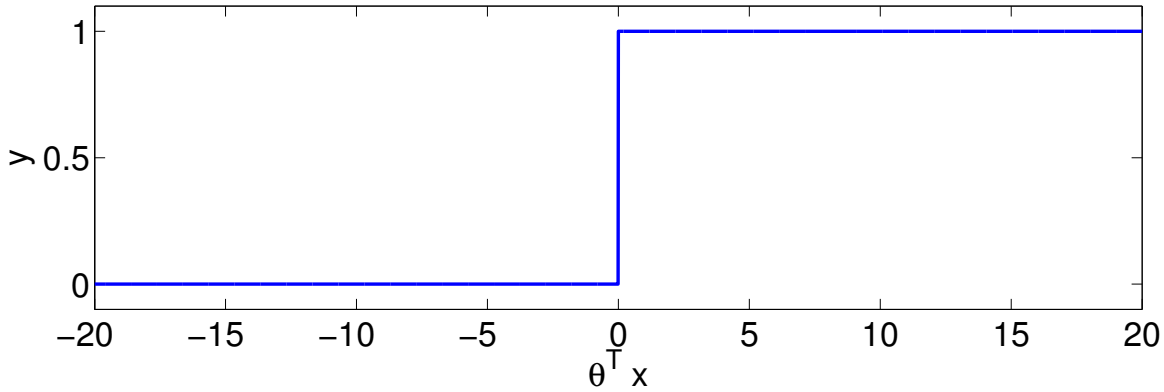
Rule

$$y = \begin{cases} 0 & \text{if } h_{\boldsymbol{\theta}}(\mathbf{x}) < 1/2 \\ 1 & \text{if } h_{\boldsymbol{\theta}}(\mathbf{x}) \geq 1/2 \end{cases}$$

Equivalent rule

$$y = \begin{cases} 0 & \text{if } \boldsymbol{\theta}^T \mathbf{x} < 0 \\ 1 & \text{if } \boldsymbol{\theta}^T \mathbf{x} \geq 0. \end{cases}$$

Prediction Rule



Rule

$$y = \begin{cases} 0 & \text{if } h_{\boldsymbol{\theta}}(\mathbf{x}) < 1/2 \\ 1 & \text{if } h_{\boldsymbol{\theta}}(\mathbf{x}) \geq 1/2 \end{cases}$$

Equivalent rule

$$y = \begin{cases} 0 & \text{if } \boldsymbol{\theta}^T \mathbf{x} < 0 \\ 1 & \text{if } \boldsymbol{\theta}^T \mathbf{x} \geq 0. \end{cases}$$

Points $\mathbf{x}$ such that $\boldsymbol{\theta}^T \mathbf{x} = 0$ compose the boundary of the classification (hyperplane !!)

The Model—Big Picture

Illustrate on board: $\mathbf{x} \rightarrow z \rightarrow p \rightarrow y$

MATLAB visualization

Cost Function

Can we used squared error?

$$J(\boldsymbol{\theta}) = \sum_i (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})^2$$

Cost Function

Can we used squared error?

$$J(\boldsymbol{\theta}) = \sum_i (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)})^2$$

This is sometimes done. But we want to do better.

Cost Function

Let's explore further. For squared error, we can write:

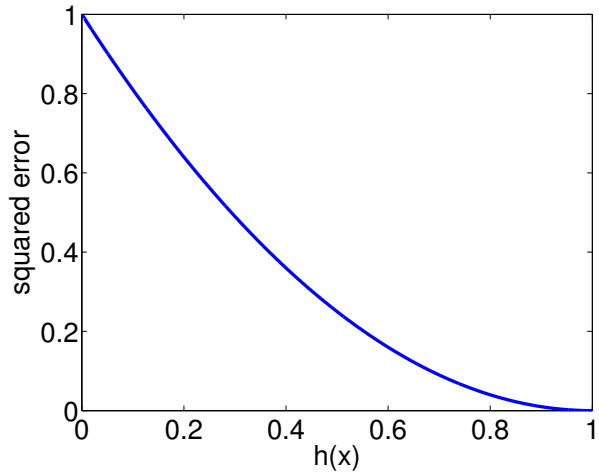
$$J(\boldsymbol{\theta}) = \sum_{i=1}^{m} \text{cost}(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}), y^{(i)})$$

$$\text{cost}(p, y) = (p - y)^2$$

$\text{cost}(p, y)$ is cost of predicting $h_{\boldsymbol{\theta}}(\mathbf{x}) = p$ when the true value is y

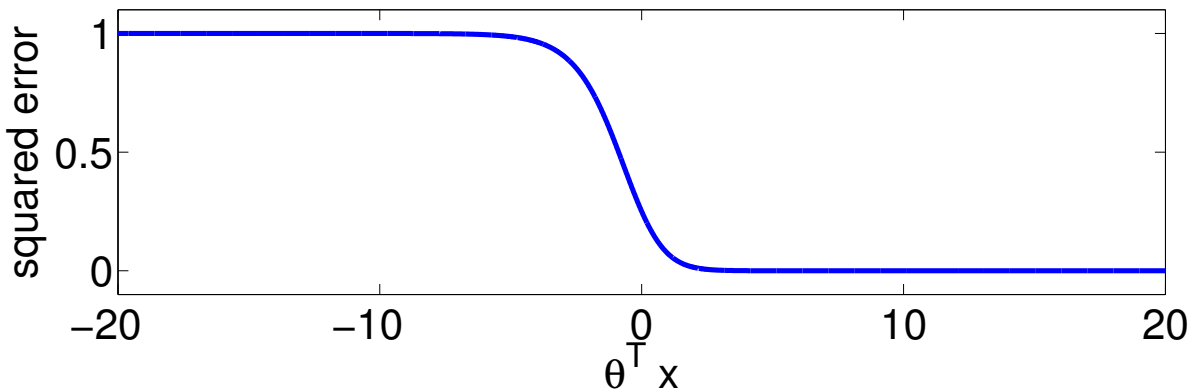
Cost Function

Suppose $y = 1$. For squared error, $\text{cost}(p, 1) = (p - 1)^2$ looks like this



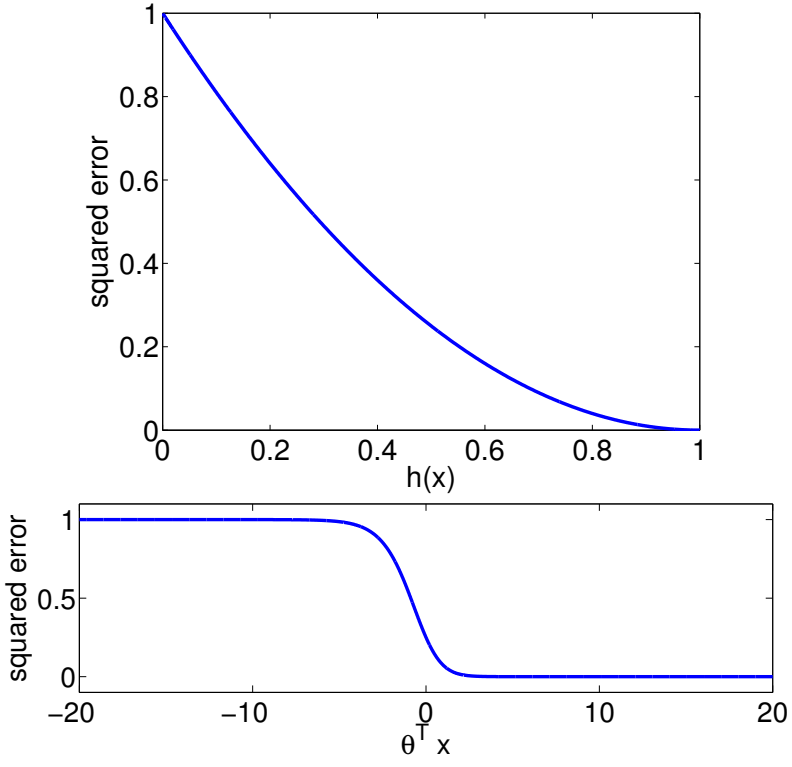
If we undo the logistic transform,

$\text{cost}(p, 1) = (h_{\boldsymbol{\theta}}(\mathbf{x}) - 1)^2 = \left(\frac{1}{1+e^{-\boldsymbol{\theta}^T \mathbf{x}}} - 1\right)^2$ looks like this



Cost Function

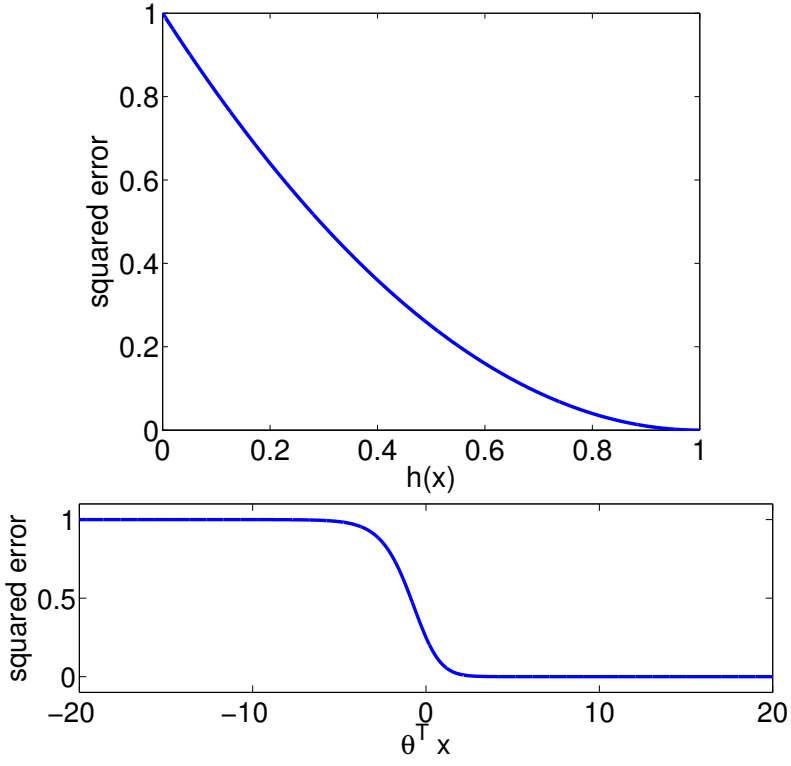
Exercise: fix these



- Recall that $y = 1$ is the correct answer

Cost Function

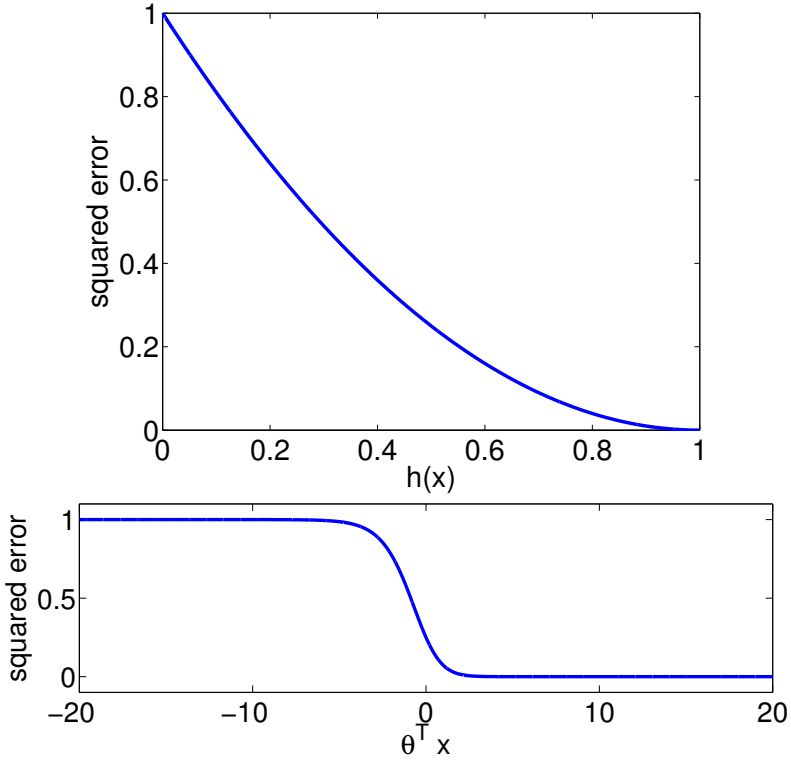
Exercise: fix these



- Recall that $y = 1$ is the correct answer
- As $z = \boldsymbol{\theta}^T \mathbf{x} \to \infty$, then $p \to 1$, so the prediction is better and better. The cost approaches zero.

Cost Function

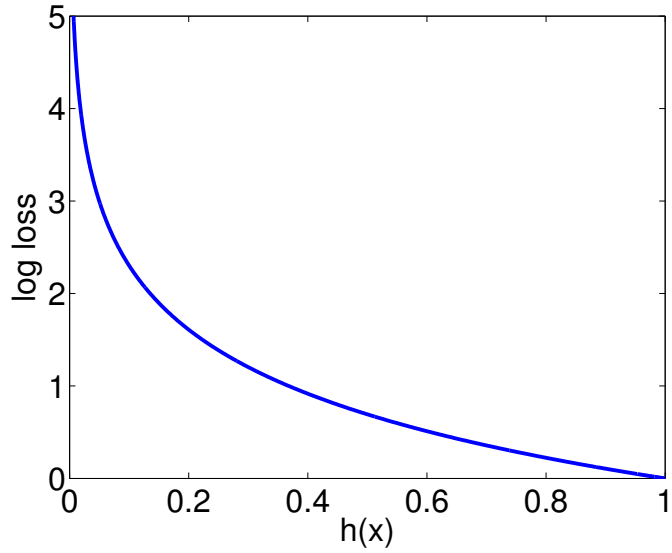
Exercise: fix these

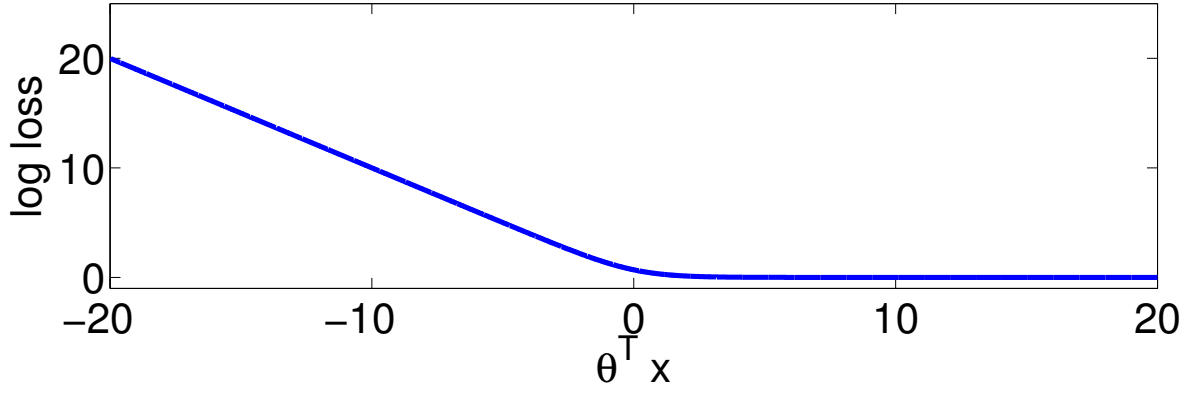


- Recall that $y = 1$ is the correct answer
- As $z = \boldsymbol{\theta}^T \mathbf{x} \to \infty$, then $p \to 1$, so the prediction is better and better. The cost approaches zero.
- As $z = \boldsymbol{\theta}^T \mathbf{x} \to -\infty$, then $p \to 0$, so the prediction is worse and worse. The cost...

Log Loss $(y = 1)$

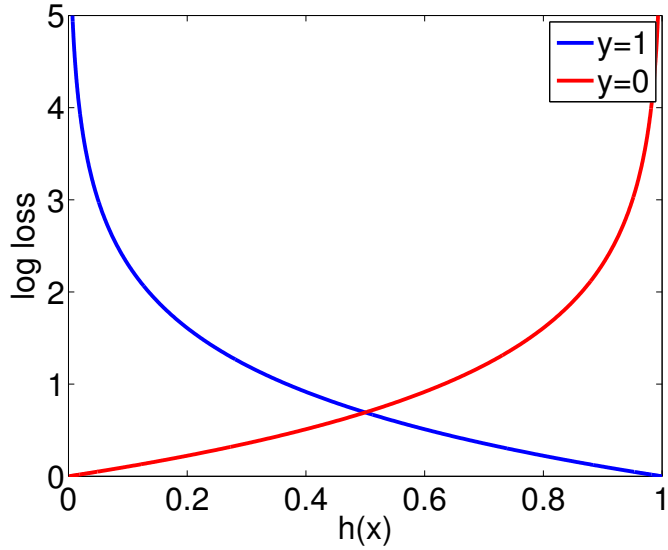
$$\text{cost}(p, 1) = -\log p$$

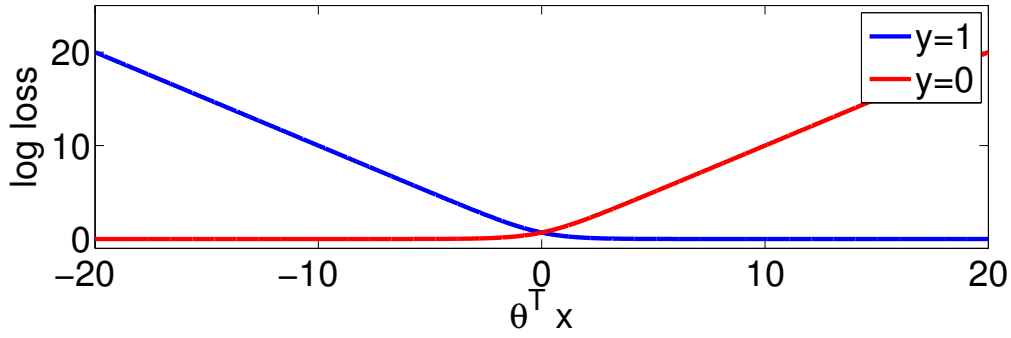




Log Loss

$$\text{cost}(p, y) = \begin{cases} -\log p & y = 1 \\ -\log(1-p) & y = 0 \end{cases}$$





Equivalent Expression for Log-Loss

$$\mathsf{cost}(p, y) = \begin{cases} -\log p & y = 1 \\ -\log(1-p) & y = 0 \end{cases}$$

$$\mathsf{cost}(p, y) = -y \log p - (1-y) \log(1-p)$$

Equivalent Expression for Log-Loss

$$\mathsf{cost}(p, y) = \begin{cases} -\log p & y = 1 \\ -\log(1-p) & y = 0 \end{cases}$$

$$\mathsf{cost}(p, y) = -y \log p - (1-y) \log(1-p)$$

$$\mathsf{cost}(h_{\boldsymbol{\theta}}(\mathbf{x}), y) = -y \log h_{\boldsymbol{\theta}}(\mathbf{x}) - (1-y) \log(1 - h_{\boldsymbol{\theta}}(\mathbf{x}))$$

Review so far

- Input: $\mathbf{x} \in \mathbb{R}^n$
- Output: $y \in \{0, 1\}$
- Model (hypothesis class)

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \text{logistic}(\boldsymbol{\theta}^T \mathbf{x}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}}}$$

- Cost function:

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{m} \Big(- y^{(i)} \log h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - (1 - y^{(i)}) \log(1 - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})) \Big)$$

TODO: optimize $J(\boldsymbol{\theta})$

Gradient Descent for Logistic Regression

1. Initialize $\theta_0, \theta_1, \ldots, \theta_d$ arbitrarily
2. Repeat until convergence

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}), \qquad j = 0, \ldots, d.$$

Gradient Descent for Logistic Regression

1. Initialize $\theta_0, \theta_1, \ldots, \theta_d$ arbitrarily
2. Repeat until convergence

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}), \qquad j = 0, \ldots, d.$$

Partial derivatives for logistic regression (exercise):

$$\frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta}) = \sum_{i=1}^{m} (h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) - y^{(i)}) x_j^{(i)}$$

(Same as linear regression! But $h_{\boldsymbol{\theta}}(\mathbf{x})$ is different)

Decision Boundaries

Example from R&N (Fig. 18.15).

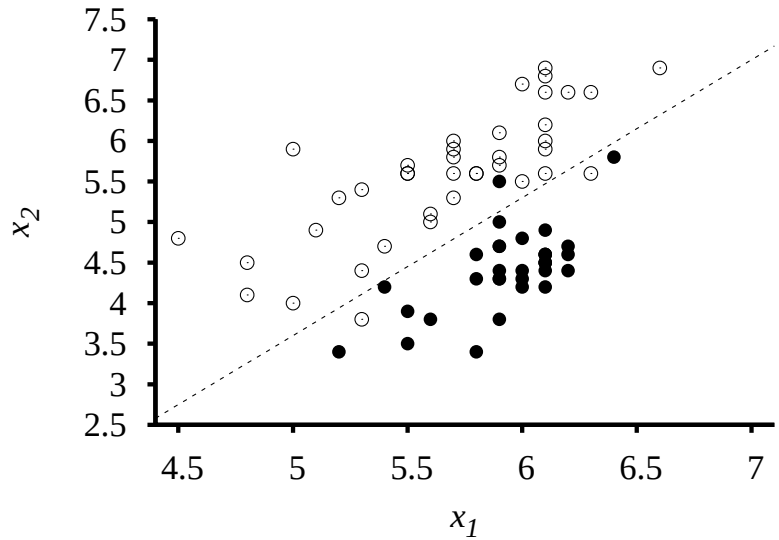
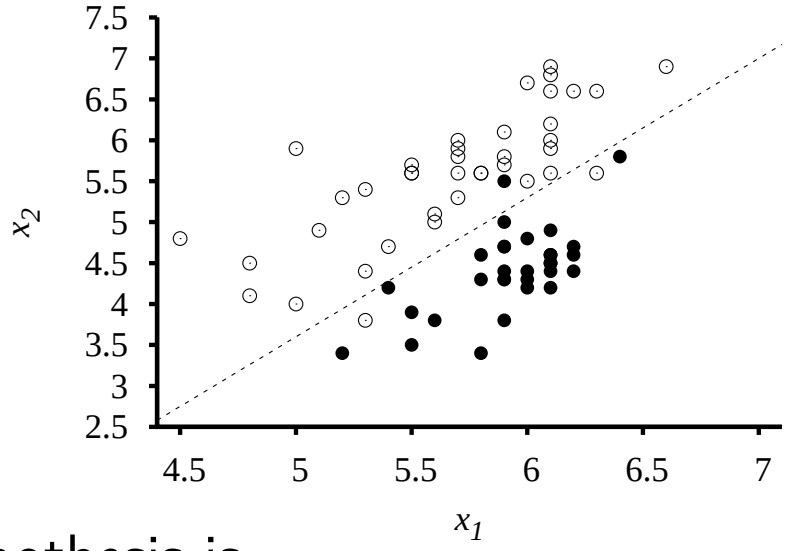


Figure 1: Earthquakes (white circles) vs. nuclear explosions (black circles) by body wave magnitude ($x1$) and surface wave magnitude ($x2$)

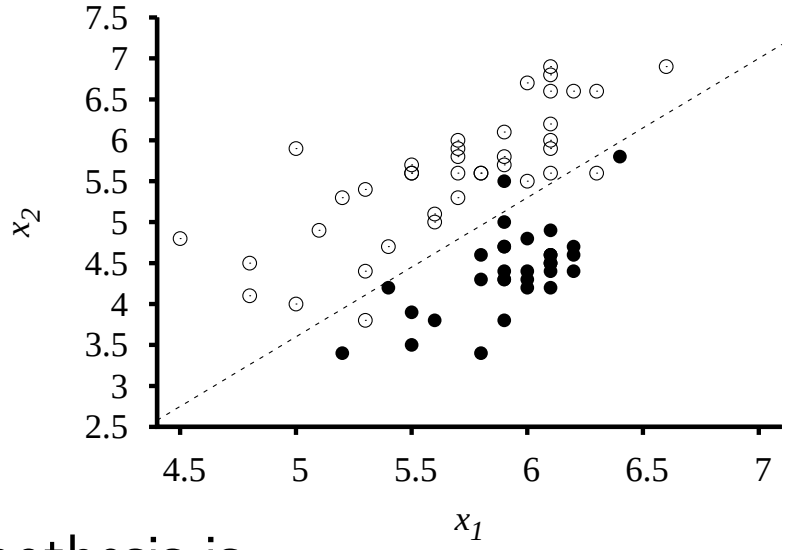
Decision Boundaries



E.g., suppose hypothesis is

$$h(x_1, x_2) = \text{logistic}(1.7x_1 - x_2 - 4.9)$$

Decision Boundaries



E.g., suppose hypothesis is

$$h(x_1, x_2) = \text{logistic}(1.7x_1 - x_2 - 4.9)$$

Predict nuclear explosion if:

$$1.7x_1 - x_2 - 4.9 \geq 0$$

$$x_2 \leq 1.7x_1 - 4.9$$

Linear Classifiers

Predict

$$y = \begin{cases} 0 & \text{if } \boldsymbol{\theta}^T \mathbf{x} < 0, \\ 1 & \text{if } \boldsymbol{\theta}^T \mathbf{x} \geq 0. \end{cases}$$

Watch out! Hyperplane!

Many other learning algorithms use linear classification rules

- Perceptron
- Support vector machines (SVMs)
- Linear discriminants

Nonlinear Decision Boundaries by Feature Expansion

Example (Ng)

$$(x_1, x_2) \mapsto (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2),$$

$$\boldsymbol{\theta} = \begin{bmatrix} -1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}^T$$

Exercise: what does decision boundary look like in (x_1, x_2) plane?

Nonlinear Decision Boundaries by Feature Expansion

Example (Ng)

$$(x_1, x_2) \mapsto (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2),$$

$$\boldsymbol{\theta} = \begin{bmatrix} -1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}^T$$

Exercise: what does decision boundary look like in (x_1, x_2) plane?

$$\boldsymbol{\theta}^T \mathbf{x} = -1 + x_1^2 + x_2^2 = 0$$

Nonlinear Decision Boundaries by Feature Expansion

Example (Ng)

$$(x_1, x_2) \mapsto (1, x_1, x_2, x_1^2, x_2^2, x_1 x_2),$$

$$\boldsymbol{\theta} = \begin{bmatrix} -1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}^T$$

Exercise: what does decision boundary look like in (x_1, x_2) plane?

$$\boldsymbol{\theta}^T \mathbf{x} = -1 + x_1^2 + x_2^2 = 0$$

$$x_1^2 + x_2^2 \leq 1 \quad \text{class0}$$
$$x_1^2 + x_2^2 > 1 \quad \text{class1}$$

Note: Where Does Log Loss Come From?

Arises naturally from the probabilistic interpretation of logistic regression