Applied
Machine Learning

Lecture: 5

Logistic Regression

Ekarat Rattagan, Ph.D.

Slides adapted from Andrew NG, Eric Eaton, Raquel Urtasun, and Patrick Winston

Outline

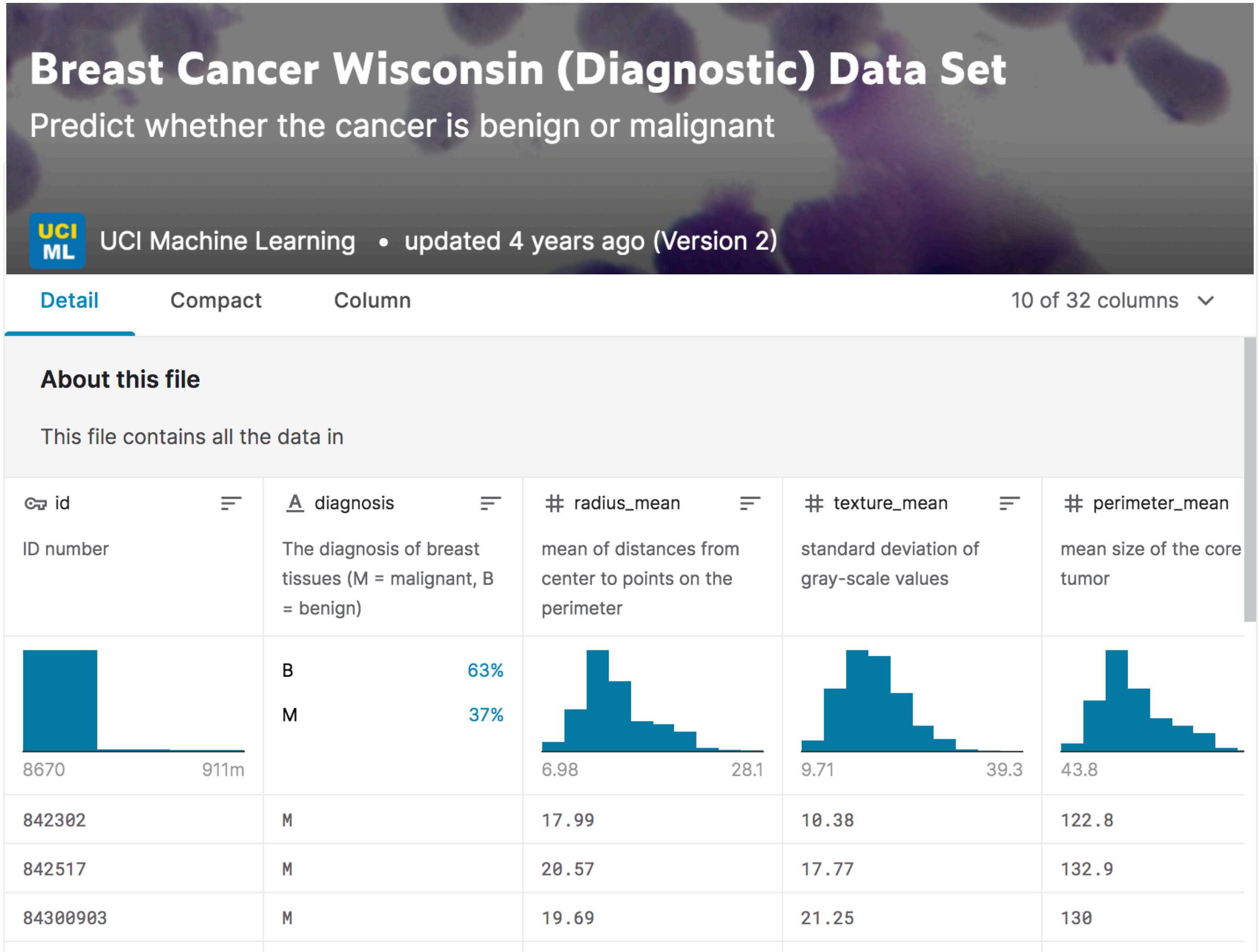
5.1 Regression VS Classification
5.2 Logistic Regression
5.3 Decision boundary
5.4 Cost function
5.5 Gradient descent

5.1 Regression VS Classification

Classification VS Regression

- Supervised ML is interested in mapping the input x to a label y
- In regression —> $y \in \mathbb{R}$
    - House price prediction
- In classification —> y is categorical, e.g., $y \in \{0, 1\}$
    - Email: Spam / Not Spam?
    - Online Transactions: Fraudulent (Yes / No)?
    - Tumor: Malignant / Benign ?

Classification problem

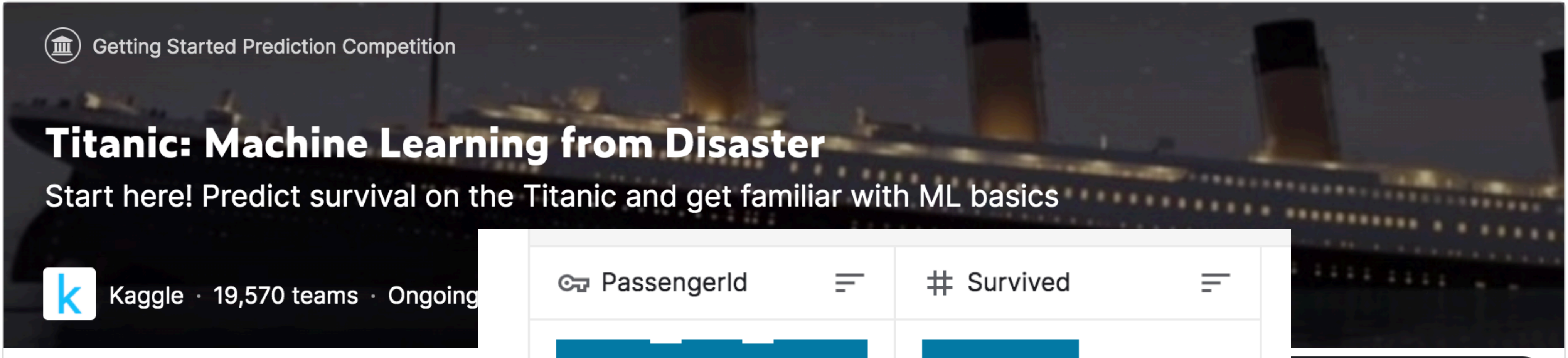


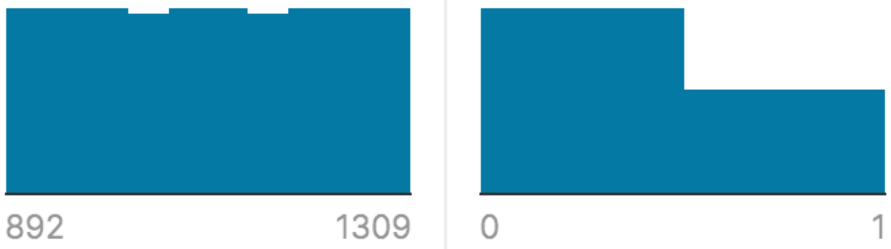
https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Tumor: Malignant / Benign ?

$$y \in \{0, 1\}$$

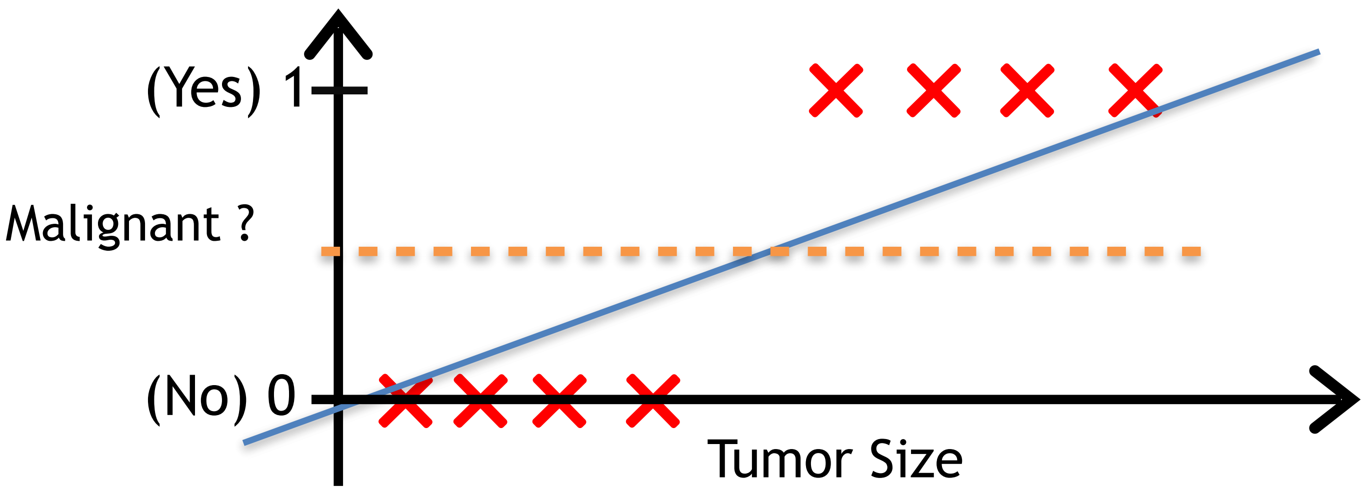
0: “Negative Class” (e.g., benign tumor)
1: “Positive Class” (e.g., malignant tumor)





PassengerId	Survived
892	0
893	1
894	0
895	0

https://www.kaggle.com/c/titanic/



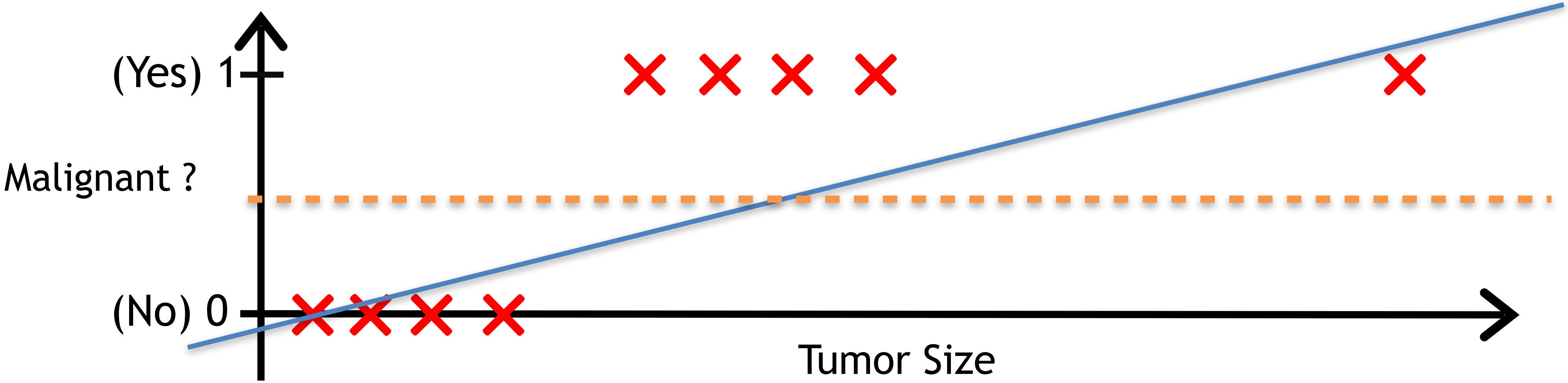
A reasonable decision rule

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict “y = 1”

If $h_\theta(x) < 0.5$, predict “y = 0”

Credit: Andrew NG



A reasonable decision rule (How can I mathematically write this rule?)

Threshold classifier output $h_\theta(x)$ at 0.5:

If $h_\theta(x) \geq 0.5$, predict "y = 1"

If $h_\theta(x) < 0.5$, predict "y = 0"

Credit: Andrew NG

Classification: $y = 0$ or 1

Linear Regression:

$h_\theta(x)$ can be < 0 or > 1

Logistic Regression:

$$0 \leq h_\theta(x) \leq 1$$

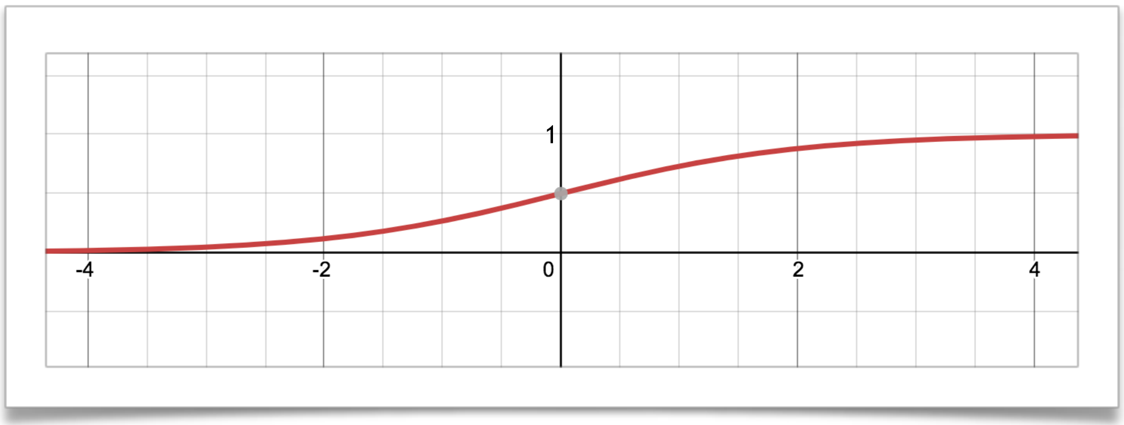
Credit: Andrew NG

5.2 Logistic Regression

Logistic Regression Model

We applied **sigmoid function** to a linear function of the data

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




```python
def sigmoid(z):
    """ sigmoid """
    return 1 / (1 + np.exp(-z))
```

Logistic Regression Model

We applied **sigmoid function** to a linear function of the data

$$h_\theta(x) = \theta^T x$$

$$h_\theta(x) = \sigma(\theta^T x)$$

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

Interpretation of Hypothesis Output

$h_\theta(x)$ = estimated probability that $y = 1$ on input x

$h_\theta(x) = 0.7$ tell patient that 70% chance of tumor being malignant

$$P(y = 0|x;\theta) + P(y = 1|x;\theta) = 1$$

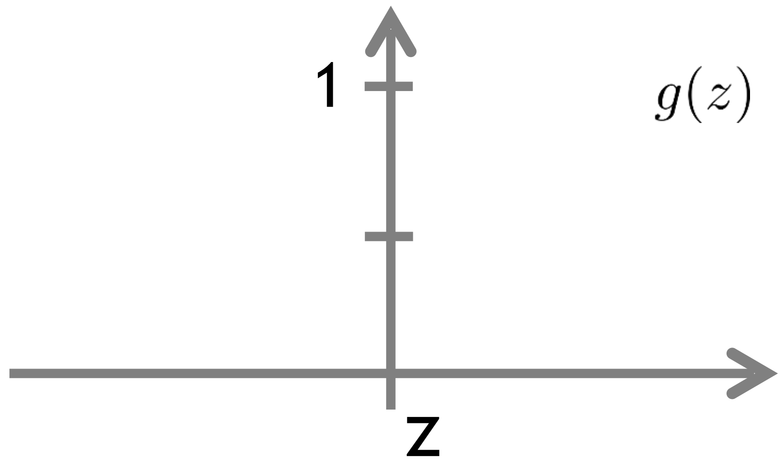
Probability that y = 1, given x, parameterized by θ

Probability that y = 0, given x, parameterized by θ

$$P(y = 0|x;\theta) = 1 - P(y = 1|x;\theta)$$

5.3 Decision boundary

Logistic regression



$$h_\theta(x) = g(\theta^T x)$$

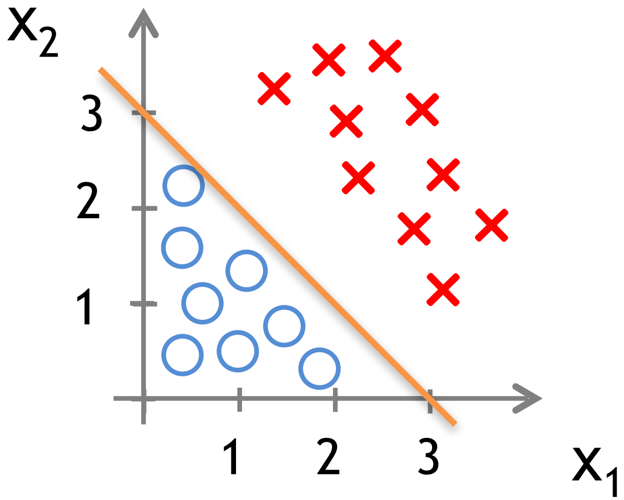
$$g(z) = \frac{1}{1+e^{-z}}$$

Suppose predict "$y = 1$" if $h_\theta(x) \geq 0.5$

predict "$y = 0$" if $h_\theta(x) < 0.5$

Credit: Andrew NG

Decision Boundary (Multiple parameters)

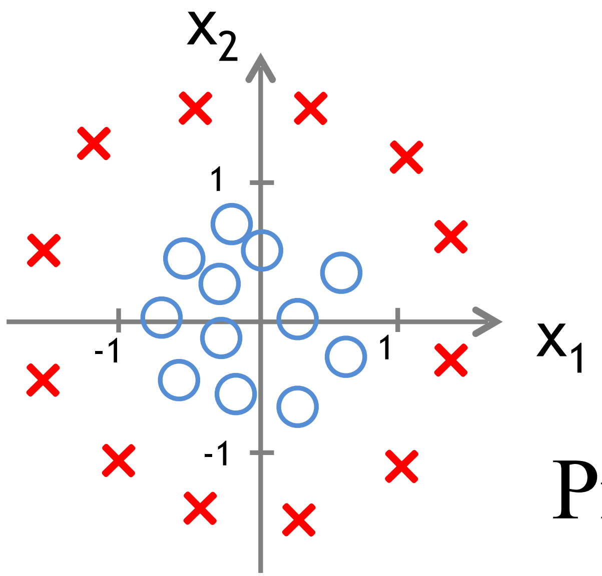


$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

Credit: Andrew NG

Non-linear decision boundaries



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \dots)$$

Credit: Andrew NG

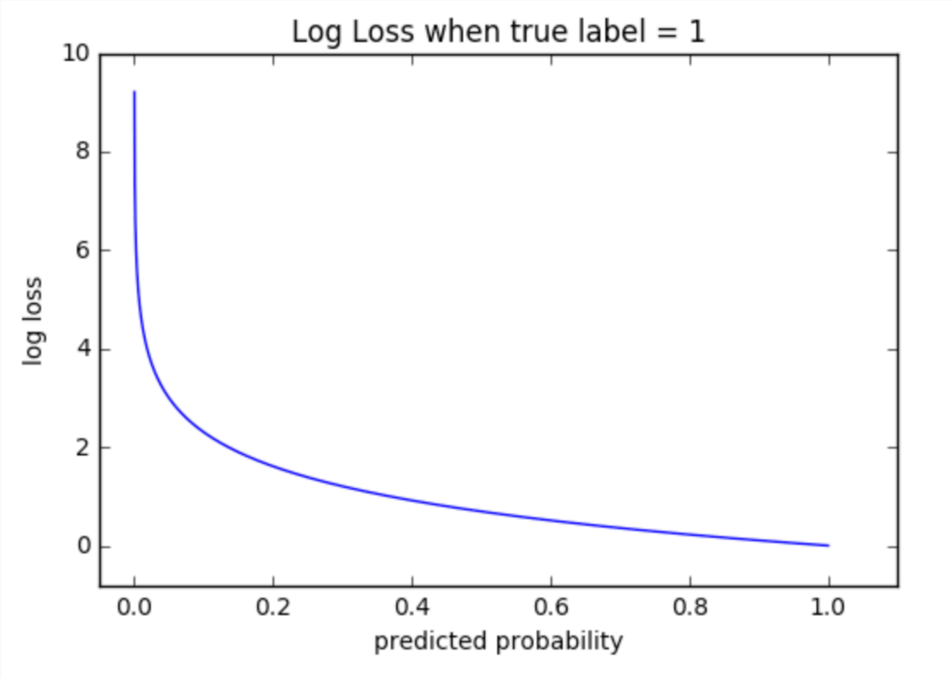
5.4 Cost function

Mean Squared Error (MSE) ?

Unfortunately we can’t use the same cost function **MSE** as we did for linear regression. Why? this is because our prediction function is non-linear (due to sigmoid transform). Squaring this prediction as we do in MSE results in **a non-convex function with many local minimums**. If our cost function has many local minimums, **gradient descent may not find the optimal global minimum.**

Cross-Entropy

Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label. So predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high loss value. A perfect model would have a log loss of 0.



The graph above shows the range of possible loss values given a true observation (isDog = 1). As the predicted probability approaches 1, log loss slowly decreases. As the predicted probability decreases, however, the log loss increases rapidly. Log loss penalizes both types of errors, but especially those predictions that are confident and wrong!

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Cost}(h_\theta(x^{(i)}), y^{(i)})$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(h_\theta(x)) \qquad \text{if y = 1}$$

$$\mathrm{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \qquad \text{if y = 0}$$

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

Bernoulli distribution

$$p(y^{(i)} = 1 \mid x^{(i)}; \theta) = h_\theta(x^{(i)})$$

$$+$$

$$p(y^{(i)} = 0 \mid x^{(i)}; \theta) = 1 - h_\theta(x^{(i)})$$

$$p(y^{(i)} \mid x^{(i)}; \theta) = h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

5.5 Gradient descent

Partial derivative

$$J(\theta) \quad = \quad -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log(h_\theta(x^{(i)})) + (1 - y^{(i)})log(1 - h_\theta(x^{(i)}))\right]$$

$$\frac{\partial J(\theta)}{\partial \theta} \quad = \quad -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{h_\theta(x^{(i)})}\cdot\frac{\partial h_\theta(x^{(i)})}{\partial \theta} \quad + \quad \frac{(1 - y^{(i)})}{(1 - h_\theta(x^{(i)}))}\cdot\ (-1)\cdot\frac{\partial h_\theta(x^{(i)})}{\partial \theta}\right]$$

Partial derivative

$$\frac{\partial J(\theta)}{\partial \theta} = -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{h_\theta(x^{(i)})}\cdot\frac{\partial h_\theta(x^{(i)})}{\partial \theta} + \frac{(1-y^{(i)})}{(1-h_\theta(x^{(i)}))}\cdot(-1)\cdot\frac{\partial h_\theta(x^{(i)})}{\partial \theta}\right]$$

$$\frac{\partial h_\theta(x^{(i)})}{\partial \theta} = \frac{\partial \sigma(\theta^T x)}{\partial \theta} = \frac{\partial \sigma(\theta^T x)}{\partial(\theta^T x)}\cdot\frac{\partial \theta^T x}{\partial \theta}$$

$$x_j^i$$

$$\sigma(\theta^T x)\cdot(1-\sigma(\theta^T x))$$

$$h_\theta(x^{(i)})\cdot(1-h_\theta(x^{(i)}))\ \ x_j^i$$

Partial derivative

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{h_\theta(x^{(i)})}\cdot\frac{\partial h_\theta(x^{(i)})}{\partial \theta} + \frac{(1-y^{(i)})}{(1-h_\theta(x^{(i)}))}\cdot(-1)\frac{\partial h_\theta(x^{(i)})}{\partial \theta}\right]$$

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{h_\theta(x^{(i)})}\cdot h_\theta(x^{(i)})\cdot(1-h_\theta(x^{(i)}))\cdot x_j^i + \frac{(1-y^{(i)})}{(1-h_\theta(x^{(i)}))}\cdot(-1)h_\theta(x^{(i)})\cdot(1-h_\theta(x^{(i)}))\cdot x_j^i\right]$$

Partial derivative

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{-1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}}{h_\theta(x^{(i)})}\cdot h_\theta(x^{(i)})\cdot(1-h_\theta(x^{(i)}))\cdot x_j^i + \frac{(1-y^{(i)})}{(1-h_\theta(x^{(i)}))}\cdot (-1)h_\theta(x^{(i)})\cdot(1-h_\theta(x^{(i)}))\cdot x_j^i\right]$$

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{-1}{m}\sum_{i=1}^{m}\left[y^{(i)}\cdot(1-h_\theta(x^{(i)}))\cdot x_j^{(i)} + (1-y^{(i)})\cdot (-1)h_\theta(x^{(i)})\cdot x_j^{(i)}\right]$$

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{-1}{m}\sum_{i=1}^{m}\left[y^{(i)}x_j^{(i)} - y^{(i)}h_\theta(x^{(i)})x_j^{(i)} + y^{(i)}h_\theta(x^{(i)})x_j^{(i)} - h_\theta(x^{(i)})x_j^{(i)}\right]$$

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{-1}{m}\sum_{i=1}^{m}\left[y^{(i)}x_j^{(i)} - h_\theta(x^{(i)})x_j^{(i)}\right]$$

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{-1}{m}\sum_{i=1}^{m}\left[(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\left[(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right]$$

Gradient Descent of logistic regression

Repeat until convergence $\{$

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left[(h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \right]$$

$\}$