Logistic Regression

Imagine you have credit card transaction data (time series) and you want to decide if there is fraud.

- That's an example of Binary Classification - make a yes/no decision given data.

This is probably the most common machine learning application

- Will a borrower pay back a loan?
- Will it rain tomorrow?
- Is this an image of a cat?

Logistic Regression

Given data $\mathbf{x}$, a program that simply outputs a binary 0 - 1/yes - no response isn't very informative. Ideally, would like

$$P(Y = 1 \mid \mathbf{x})$$

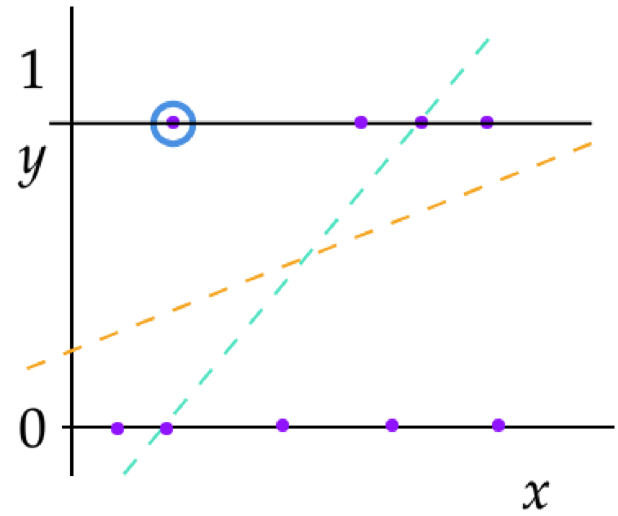
This is the probability that the response is (yes).

Why not use linear regression?

$$y = b + \mathbf{w} \cdot \mathbf{x}$$

Several problems:

- The computed y can be < 0 and > 1
- So definitely not a probability.
- Strongly influenced by outliers



Logistic Regression

Assume that there is data $\{\mathbf{x}^i, y^i\}$ where the $\mathbf{x}^i$are data/measurement vectors and the y^i are the associated 0-1 responses.

Let

$$p(\mathbf{x}^i, \theta) = p(y^i = 1|\mathbf{x}^i)$$

Assuming the samples $\mathbf{x}^i, y^i$ are independent, the likelihood function is:

$$\prod_{i=1}^{m} p(y^i|\mathbf{x}^i) = \prod_{i}^{m} p(\mathbf{x}^i, \theta)^{y^i}(1 - p(\mathbf{x}^i, \theta))^{(1-y^i)}$$

That is, each sample provides data for a Bernoulli trial.

Recall, for coin tossing we had $p(\mathbf{x}^i)$ a constant so the likelihood was:

$$\prod_{i=1}^{m} p^{y^i}(1 - p)^{(1-y^i)}$$

Logistic Regression

Maximally Constrained

If p is forced to be a constant then we know setting

$$p = \frac{1}{m}\sum_{i=1}^{m} y^i$$

is optimal

This model then always predicts the most probable class independent of $\mathbf{x}^i$.

- For credit card fraud this model would be correct 99.9% of the time.
- A weather model that always predicts sunny isn't very useful.

Example

- if $p = \bar{y} = .6$ then always predict 1
- if $p = \bar{y} = .4$ then always predict 0

Logistic Regression

Minimally Constrained

On previous slide, the model $p(\mathbf{x}^i, \theta)$ was constrained to be the same for all $\mathbf{x}^i$. What if $p(\mathbf{x}^i; \theta)$ is allowed to change arbitrarily with each sample?

The likelihood is:

$$\prod p(\mathbf{x}^i, \theta)^{y^i}(1 - p(\mathbf{x}^i))^{1-y^i}$$

This is maximized by taking $p(\mathbf{x}^i, \theta) = y^i$

A model that just returns the known outcomes is again not useful.

- **There needs to be a constraint on $p(\mathbf{x}^i, \theta)$, but it needs to be looser than forcing it to be a constant.**

Logistic Regression

What if $p(\mathbf{x}^i, \theta)$ is constrained to be a linear function of $\mathbf{x}$?

- Not restricted to $[0, 1]$
- Doubling $\mathbf{x}$ always doubles p which is not always reasonable

What if $\ln(p(\mathbf{x}^i, \theta))$ is linear in $\mathbf{x}$

- $p = e^{b+\mathbf{w}\cdot\mathbf{x}} > 0$ but not <= 1

- This gives needed restriction on p as $\mathbf{x} \to -\infty$ but not as $\mathbf{x} \to \infty$ (assuming $\mathbf{w}$ is positive)

p must have an upper bound of 1, but not the odds function. The **odds-to-be-one** function is:

$$\frac{p}{1-p} \in [0, \infty)$$

So, $\frac{p}{1-p} \sim e^{b+\mathbf{w}\cdot\mathbf{x}}$ makes sense

Logistic Regression

Logistic Model:

$$\begin{aligned}
\ln \frac{p(\mathbf{x})}{1 - p(\mathbf{x}} &= b + \mathbf{w} \cdot \mathbf{x} \\
p(\mathbf{x};\ b, \mathbf{w}) &= \frac{e^{b+\mathbf{w}\cdot\mathbf{x}}}{1 + e^{b+\mathbf{w}\cdot\mathbf{x}}} \\
&= \frac{1}{1 + e^{-(b+\mathbf{w}\cdot\mathbf{x})}} \\
&= \sigma(b + \mathbf{w} \cdot \mathbf{x})
\end{aligned}$$

That is, the log-odds are given by the sigmoid function σ with parameters b and $\mathbf{w}$

Sigmoid (Logistic) Function

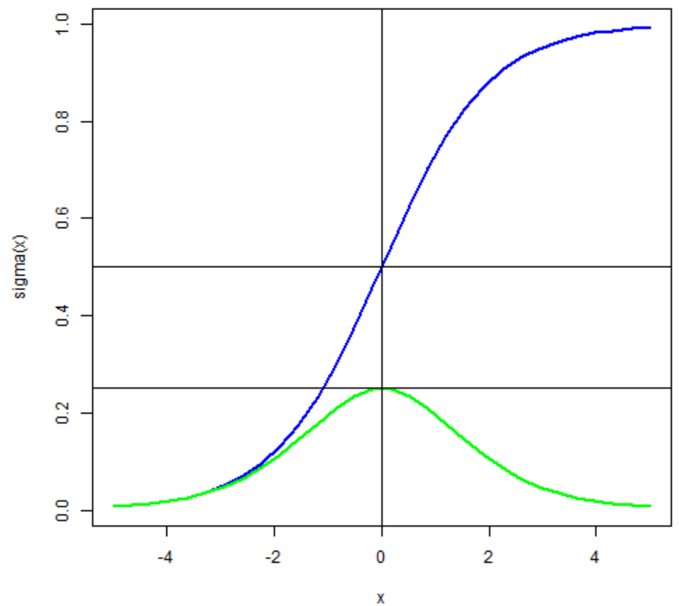
Standard Sigmoid Function

$$\sigma(x) = \frac{1}{1+e^{-x}}, \quad \sigma(0) = \frac{1}{2}$$

$$\sigma' = \sigma(1-\sigma), \quad \sigma'(0) = \frac{1}{4}$$

$$\lim_{x \uparrow +\infty} \sigma(x) = 1, \lim_{x \downarrow -\infty} \sigma(x) = 0$$

$$\lim_{x \uparrow +\infty} \sigma'(x) = 0, \lim_{x \downarrow -\infty} \sigma'(x) = 0$$



The sigmoid function is the CDF of the logistic distribution

Logistic Regression

Logistic Decision Boundary

Given estimates for $b, \mathbf{w}$ (covered later) the model predicts 1 (yes) if

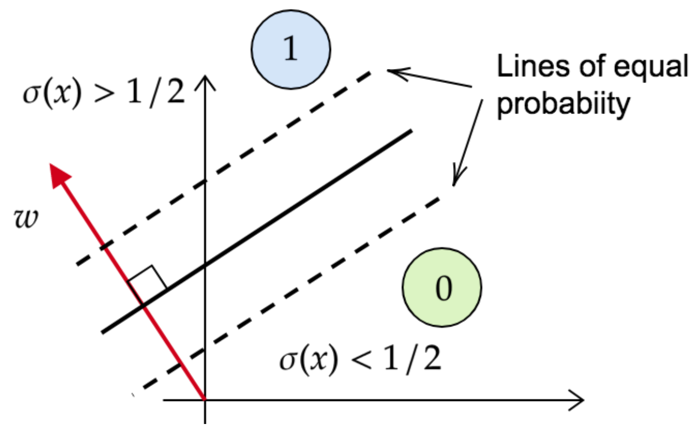
$$p(\mathbf{x}) = \sigma(b + \mathbf{w} \cdot \mathbf{x}) \geq \frac{1}{2}$$

The estimated probability will be constant along lines where $b + \mathbf{w} \cdot \mathbf{x}$ is constant.

Since the $\sigma(0) = \frac{1}{2}$, the logistic decision rule is

$$y_{pred} = \begin{cases} 1 & \text{if} \quad b + \mathbf{w} \cdot \mathbf{x} \geq 0 \\ 0 & \qquad\qquad\qquad \text{o.w.} \end{cases}$$

Decision boundary is $z = b + \mathbf{w} \cdot \mathbf{x} = 0$



Logistic Regression

Logistic regression is the workhorse of binary classification

- Commonly used as a benchmark for neural network models

Warning!

- When the data is separable, training will make the probability gradient steeper and steeper perpendicular to the decision boundary
- Can appear that it isn't converging
- This effect is usually offset by regularizing the model parameters. Will cover regularization later in course.

Logistic Regression Justification

Logistic regression models the log-odds as a linear function of the data $\mathbf{x}$. This holds if the underlying class-conditional distributions are Gaussian with shared covariance:

Log odds

$$logit(p(c_1|\mathbf{x})) = \ln \frac{p(c_1|\mathbf{x})}{1 - p(c_1|\mathbf{x})} = \ln \frac{p(c_1|\mathbf{x})}{p(c_2|\mathbf{x})}$$

Apply Bayes

$$\ln \frac{p(c_1|\mathbf{x})}{p(c_2|\mathbf{x})} = \ln \frac{p(\mathbf{x}|c_1)}{p(\mathbf{x}|c_2)} + \ln \frac{p(c_1)}{p(c_2)}$$

Now assume the class conditional probabilities are multidimensional Gaussian ($\mathbf{x} \in \mathbb{R}^d$)

$$p(\mathbf{x}|c_i) \sim \frac{1}{(2\pi)^{\frac{-d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma^{-1} (\mathbf{x}-\mu_i)}$$

Note: Σ is the covariance matrix, $|\Sigma|$ is the determinant

Logistic Regression Justification

Some algebra gives linear relation for log-odds.

$$\ln \frac{p(c_1|\mathbf{x})}{1 - p(c_1|\mathbf{x})} = \mathbf{w}^T\mathbf{x} + w_0$$

where

$$\begin{aligned} \mathbf{w} &= \Sigma^{-1}(\mu_1 - \mu_2) \\ w_0 &= -\frac{1}{2}(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2) + \ln \frac{p(c_1)}{p(c_2} \end{aligned}$$

Inverting the log-odds gives

$$\begin{aligned} p(c_1|\mathbf{x}) &= \sigma(w_0 + \mathbf{w}^T\mathbf{x}) \\ &= \frac{1}{1 + e^{-(w_0 + \mathbf{w}^T\mathbf{x})}} \end{aligned}$$

Logistic Regression - Need for Proxy Cost Function

Prediction Error Cost Function

First idea:

Initialize $b, \mathbf{w}$

Compute
$p^{(i)} = p(\mathbf{x}^i, b, \mathbf{w})$ for $i = 1, \cdots, m$

Compute predictions

$$y^i_{\text{pred}} = \begin{cases} 1 & \text{if} \quad p^i \geq \frac{1}{2} \\ 0 & \text{o.w.} \end{cases}$$

Do gradient descent on

$$C = \frac{1}{m} \sum_i |y^i_{\text{pred}} - y^i|$$

Problem:

- Small changes in $b, \mathbf{w}$ won't affect y^i_{pred} unless $b + \mathbf{w} \cdot \mathbf{x^i} = \mathbf{0}$
- Gradient steps will not receive 'feedback' from discrete cost function

Logistic Regression

There are many model choices, logistic is just one. Modeling is always about making choices.

- It is unreasonable to expect data to be Gaussian with shared covariance.

- As a result, the basic assumption that the log-odds is linear in $\mathbf{x}$ is seldom justified

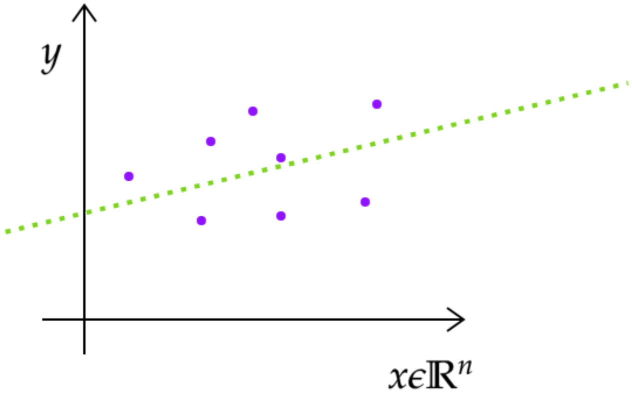
- Surprisingly, logistic regression often works very well as a binary classifier.

Least Squares Regression

Have m data points $\mathbf{x}^{(i)} \in \mathbb{R}^n$ with corresponding scalar response values $y^{(i)}$. The data samples are arranged in rows in the matrix X. X is $m \times n$

$$X = \begin{bmatrix} \mathbf{x}_1^{(1)} & \cdots & \mathbf{x}_n^{(1)} \\ & \vdots & \vdots \\ \mathbf{x}_1^{(m)} & \cdots & \mathbf{x}_n^{(m)} \end{bmatrix}$$

$$Y = \begin{bmatrix} y^{(i)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

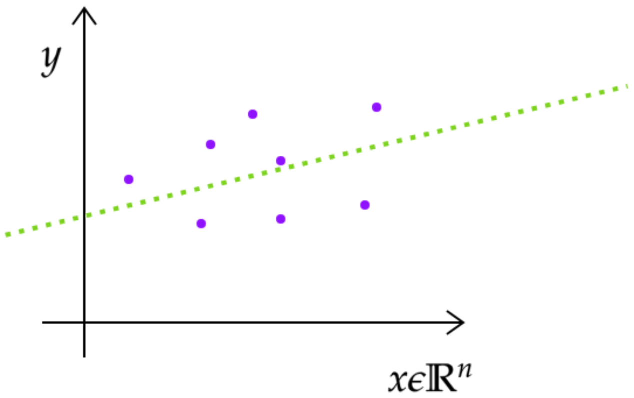


Least Squares Regression

Solution via Projection

Previously showed that the projection of y onto the subspace of $\mathbb{R}^m$ spanned by the columns of X (assuming independent cols) was

$$y_{\text{proj}} = (X^T X)^{-1} X^T Y$$



Notes:

1. Arranged samples in rows so covariance is $X^T X$
2. Assume $\mathbf{x}^i = 1$ to allow for intercept

Least Squares Regression

The traditional (calculus) approach is to assume a linear relation between $y^{(i)}$ and $x^{(i)}$

$$\hat{y}^{(i)} = w_0 + w_1 \mathbf{x}_1^{(i)} + \cdots + w_n \mathbf{x}_n^{(i)}$$

So the error in the estimate is:

$$e^{(i)} = \hat{y}^{(i)} - \mathbf{w}^T \cdot (1, \mathbf{x}_1^{(i)}, \cdots, \mathbf{x}_n^{(i)})$$

In vector form

$$\mathbf{e} = \mathbf{y} - \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_n^{(1)} \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_n \end{bmatrix}$$

Below,the tilde over X indicates that a column of 1s has been added to the data matrix X. Want to minimize cost $C(w)$.

$$\mathbf{e} = \mathbf{y} - \tilde{X}\mathbf{w}$$

$$\|\mathbf{e}\|^2 = \sum_i (y^{(i)} - \mathbf{w} \cdot (1, \mathbf{x}_1^{(i)}, \cdots, \mathbf{x}_n^{(i)}))^2$$

Least Squares Regression

Traditional Calculus Approach

The cost function is the magnitude of the error vector

$$
\begin{aligned}
\mathcal{C}(\mathbf{w}) &= \sum_i (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 \\
&= \|\mathbf{e}\|^2 \\
&= \mathbf{e}^T \mathbf{e} \\
&= (\mathbf{y} - \tilde{X}\mathbf{w})^T (\mathbf{y} - \tilde{X}\mathbf{w}) \\
&= \|\mathbf{y}\|^2 - 2\mathbf{y}^T \tilde{X}\mathbf{w} + \mathbf{w}^T \tilde{X}^T \tilde{X}\mathbf{w}
\end{aligned}
$$

$$\frac{\partial C}{\partial \mathbf{w}} = -2\tilde{X}^T \mathbf{y} + 2\tilde{X}^T \tilde{X}\mathbf{w}$$

setting

$$\frac{\partial C}{\partial \mathbf{w}} = 0$$

gives

$$\tilde{X}^T \tilde{X}\mathbf{w} = \tilde{X}^T \mathbf{y}$$

$$\mathbf{w} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \mathbf{y}$$

showing

$$\mathbf{w} = \hat{Y}_{proj}$$

There is a third way: Maximum Likelihood

Least Squares Regression

Maximum Likelihood

Again, assume the response variable $y^{(i)}$ is linearly related to a weight vector.

$$\mathbf{y}^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$$

$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

Will show that if the $\epsilon^{(i)}$ are iid, zero mean with shared variance σ, then the parameter estimates $(\mathbf{w})$ match earlier results.

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\epsilon^i)^2}{2\sigma^2}}$$

substituting for $\epsilon^{(i)}$ gives

$$p(\mathbf{y}^{(i)} | \mathbf{x}^{(i)}; \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mathbf{y}^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2}{2\sigma^2}}$$

Least Squares Regression

The likelihood of iid data is the product of the indivdual event probabilities

$$
\begin{aligned}
L(\mathbf{w}) &= \prod_{i=1}^{m} p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)};\mathbf{w}) \\
l(\mathbf{w}) &= \ln(L(\mathbf{w})) \\
&= \ln \prod_{i=1}^{m} p^{(i)} \\
&= \sum \ln\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\mathbf{y}^{(i)}-\mathbf{w}^T\mathbf{x}^{(i)})^2}{2\sigma^2}}\right) \\
&= m \ln \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (\mathbf{y}^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2
\end{aligned}
$$

The first term above doesn't depend on $\mathbf{w}$, so maximizing $l(\mathbf{w})$ is identical to minimizing the least squares error function

$$\sum_{i=1}^{m} (\mathbf{y}^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2$$

Least Squares Regression - Convexity

Want to show that $\|\mathbf{y} - X\mathbf{w}\|^2$ is convex in $\mathbf{w}$

Method 1

First, the L_2 norm $\|\cdot\|_2$ is convex. For $t \in (0, 1)$

$$\begin{aligned} \|t\mathbf{x} + (1-t)\mathbf{y})\| &\leq \|t\mathbf{x}\| + \|(1-t)\mathbf{y})\| && \text{triangle inequality} \\ &= t\|\mathbf{x}\| + (1-t)\|\mathbf{y}\| && t, (1-t) \geq 0 \end{aligned}$$

Next, the composition of a convex function and an affine function is convex. If $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then $f(A\mathbf{x} + b)$ is convex in $\mathbf{x}$.

$$\begin{aligned} g(\mathbf{x}) &= A\mathbf{x} + b \\ g(t \cdot \mathbf{x} + (1-t) \cdot \mathbf{y}) &= A(t \cdot \mathbf{x} + (1-t)\mathbf{y}) + b \\ &= A(t \cdot \mathbf{x} + (1-t)\mathbf{y}) + tb + (1-t)b \\ &= t(A\mathbf{x} + b) + (1-t)(A\mathbf{y} + b) \\ &= tg(\mathbf{x}) + (1-t)g(\mathbf{y}) \end{aligned}$$

So, $g(\mathbf{x})$ is convex.

Least Squares Convexity

Let f be convex and consider composition $f \circ g$

$$
\begin{aligned}
f(g(t\mathbf{x} + (1-t)\mathbf{y}) &= f(tg(\mathbf{x}) + (1-t)g(\mathbf{y})) && \text{by convexity of } g \\
&\leq tf(g(\mathbf{x})) + (1-t)f(g(\mathbf{y})) && \text{by convexity of } f
\end{aligned}
$$

So, composition is convex in $\mathbf{x}$.

Together, these results show that the composition $\|\mathbf{y} - X\mathbf{w}\|$ is convex in $\mathbf{w}$
Finally, $\|\mathbf{y} - X\mathbf{w}\|^2$ is composition of a convex function with a non-decreasing convex function $\mathbf{y} = \mathbf{x}^2$, so convex.

Least Squares Convexity

Method 2

Show the Hessian is positive semi-definite.

$$\begin{aligned}
C(\mathbf{w}) &= \|\mathbf{y} - X\mathbf{w}\|^2 \\
\frac{\partial C}{\partial \mathbf{w}} &= -2\mathbf{y}^T X + 2(X\mathbf{w}^T)X \\
&= -2\mathbf{y}^T X + 2\mathbf{w}^T X^T X \\
\frac{\partial^2 C}{\partial \mathbf{w}^2} &= 2X^T X
\end{aligned}$$

$X^T X$ is positive semi-definite because

$$\mathbf{y}^T X^T X \mathbf{y} = (X\mathbf{y})^T (X\mathbf{y}) = \|X\mathbf{y}\|^2 \geq 0$$

The Hessian will be positive definite if X has independent columns.

Least Squares Convexity

Method 3

For positive semi-definite Q, the quadratic form

$$\frac{1}{2}\mathbf{x}^T Q\mathbf{x} + c^T\mathbf{x} + b$$

is convex. It follows that the least squares cost function is convex because $X^T X$ is positive semi-definite.

$$\|\mathbf{y} - X\mathbf{w}\|^2 = \mathbf{w}^T X^T X\mathbf{w} - 2(\mathbf{y}^T X)\mathbf{w} + \mathbf{y}^T\mathbf{y}$$

Logistic Regression Convexity

Logistic (negative log-likelihood) cost function for a single sample $\mathbf{x}$, sample tag t^i and estimated probability $p(\mathbf{x}^i;\mathbf{w})$

$$C(\mathbf{w}) = -t^i \ln p(\mathbf{x}^i;\mathbf{w}) - (1-t^i)\ln(1-p(\mathbf{x}^i;\mathbf{w}))$$

Dropping sample index i

$$\begin{aligned} p(\mathbf{x},\mathbf{w}) &= \sigma(\mathbf{w}^T\mathbf{x}) \\ &= \frac{1}{1+\mathrm{e}^{-\mathbf{w}^T\mathbf{x}}} \end{aligned}$$

First compute $\frac{\partial}{\partial \mathbf{w}}(-\ln\sigma)$

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}}(-\ln\sigma) &= -\frac{1}{\sigma}\sigma(1-\sigma)\frac{\partial}{\partial \mathbf{w}}(\mathbf{w}^T\mathbf{x}) \\ &= -(1-\sigma)\mathbf{x} \end{aligned}$$

Logistic Regression Convexity

$$\frac{\partial}{\partial \mathbf{w}_i}(-\ln \sigma) = (\sigma - 1)\mathbf{x}_i$$

$$\frac{\partial^2}{\partial \mathbf{w}_j \partial \mathbf{w}_i}(-\ln \sigma) = \frac{\partial}{\partial \mathbf{w}_j}(\sigma - 1)\mathbf{x}_i$$

$$= \sigma(1-\sigma)x_i x_j$$

The Hessian of the first term of the logistic cost function for a single sample point is

$$H^i = t\sigma(1-\sigma)\mathbf{x}^i(\mathbf{x}^i)^T$$

$$H = \sum_i H^i$$

$$t \geq 0, \sigma \geq 0, 1-\sigma \geq 0$$

$$\mathbf{w}^T H \mathbf{w} = t\sigma(1-\sigma)\mathbf{w}^T\mathbf{x}\mathbf{x}^T\mathbf{w}$$

$$= t\sigma(1-\sigma)(\mathbf{x}^T\mathbf{w})^T\mathbf{x}^T\mathbf{w}$$

$$= t\sigma(1-\sigma)\|\mathbf{x}^T\mathbf{w}\| \geq 0$$

So H is positive semi-definite. Repeat process for second term to show the cost function is the positive sum of convex terms.

Gradient Descent

For consistency with later material on neural networks, will now assume that the adjustable parameters in the model are a scalar offset b and a vector $\mathbf{w}$. Given data $\mathbf{x}^{(i)}, t^{(i)}$ where $i = 1, \cdots, m$

Linear Model

$$h(X; b, \mathbf{w}) = b + X\mathbf{w}$$

Logistic Model

$$h(X; b, \mathbf{w}) = \sigma(b + X\mathbf{w})$$

For these equations X is m x n, with data vectors in rows and no constant 1 column added.

$$X = \begin{bmatrix} \mathbf{x}_1^{(1)} & \cdots & \mathbf{x}_n^{(1)} \\ \vdots & & \vdots \\ \mathbf{x}_1^{(m)} & \cdots & \mathbf{x}_n^{(m)} \end{bmatrix}$$

Note $h(X)$ is an m-element vector

Gradient Descent

The sample (training) data tag vector:

$$T = \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(m)} \end{bmatrix}$$

Where $t^{(i)} \in \mathbb{R}$ if least squares and $t^{(i)} \in (0, 1)$ if logistic regression.

Gradient Descent

Model hypothesis

$$h(\mathbf{x}) = h(\mathbf{x}; b, \mathbf{w})$$

Squared Error - h can be linear or logistic

$$\begin{aligned} C(b, \mathbf{w}) &= \frac{1}{2m} \|h(\mathbf{x}) - T\|^2 \\ &= \frac{1}{2m} \sum_{i=1}^{m} (h(\mathbf{x}^{(i)}) - t^{(i)})^2 \end{aligned}$$

Negative Log Likelihood - h logistic only

$$C(b, \mathbf{w}) = -\frac{1}{m} \sum_{i=1}^{m} \left(t^{(i)} \ln h(\mathbf{x}^{(i)}) + (1 - t^{(i)}) \ln(1 - h(\mathbf{x}^{(i)})) \right)$$

Gradient Descent

In the following, using θ as a standin for a generic adjustable parameter
Squared Error, Linear Model

$$\frac{\partial C}{\partial \theta} = \frac{1}{m}\sum_{i=1}^{m}(h(\mathbf{x}^{(i)}) - t^{(i)})\frac{\partial h^{(i)}}{\partial \theta}$$

$$\frac{\partial h}{\partial b} = 1$$

$$\frac{\partial h}{\partial \mathbf{w}} = \mathbf{x}^{(i)}$$

Letting $e^{(i)} = h(\mathbf{x}^{(i)}) - t^{(i)}$ using $e^{(i)} \in \mathbb{R}$ gives

$$\frac{\partial C}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} e^{(i)}$$

$$\frac{\partial C}{\partial \mathbf{w}} = \frac{1}{m}\sum_{i=1}^{m} e^{(i)}\mathbf{x}^{(i)}$$

Gradient Descent

Negative Log Likelihood - Logistic Model

$$C = \sum_{i=1}^{m} C^{(i)}$$

$$C^{(i)} = -t^{(i)} \ln \sigma^{(i)} - (1 - t^{(i)}) \ln(1 - \sigma^{(i)})$$

$$h^{(i)} = \sigma^{(i)} = \sigma(b + \mathbf{w} \cdot \mathbf{x}^{(i)})$$

For a single sample $\mathbf{x}^{(i)}$

$$\frac{\partial C^i}{\partial \theta} = -\frac{t^{(i)}}{h^{(i)}} \frac{\partial h^{(i)}}{\partial \theta} + \frac{1 - t^{(i)}}{1 - h^{(i)}} \frac{\partial h^{(i)}}{\partial \theta}$$
$$\frac{\partial h}{\partial \theta} = h(1 - h) \frac{\partial (b + \mathbf{w} \cdot \mathbf{x})}{\partial \theta}$$

Let $z = b + \mathbf{w} \cdot \mathbf{x}$

$$\begin{aligned}\frac{\partial C^i}{\partial \theta} &= -t^{(i)}(1 - h)\frac{\partial z}{\partial \theta} + (1 - t^{(i)})h\frac{\partial z}{\partial \theta} \\ &= (h - t^{(i)})\frac{\partial z}{\partial \theta}\end{aligned}$$

Gradient descent

For logistic, $h = \sigma$

$$\frac{\partial C^{(i)}}{\partial \theta} = (\sigma(b + \mathbf{w}^T \mathbf{x}^{(i)}) - t^{(i)}) \frac{\partial z}{\partial \theta}$$

Amazingly, get same functional form as before. Note that in least squares case $h = z$!

$$\frac{\partial C}{\partial \theta} = \sum_{i=1}^{m} C^{(i)} = \frac{1}{m} \sum_{i=1}^{m} \left(h(\mathbf{x}^{(i)}) - t^{(i)} \right) \frac{\partial z}{\partial \theta}$$

To be complete

$$\frac{\partial C}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} e^{(i)}$$

$$\frac{\partial C}{\partial \mathbf{w}} = \frac{1}{m} \sum_{i=1}^{m} e^{(i)} \mathbf{x}^{(i)}$$

Gradient Descent

Brief note on vectorizing equations

For logistic regression, the output is:

$$\sigma(b + \mathbf{w}^T \mathbf{x})$$

If the input data vectors are arranged as columns in a matrix X, then the output probabilities for all inputs can be computed using

$$\sigma(b + \mathbf{w}^T X)$$

provided σ is a vectorized function.

Gradient Descent

Select learning rate parameter α and iteratively update parameters in direction opposite to the gradient

Initialize parameters $b, \mathbf{w}$
for $i = 1, \cdots, \max iterations$ do
  compute errors $h(X; b, \mathbf{w}) - T$
  update parameters
    $b = b - \alpha \frac{\partial C}{\partial b}$
    $\mathbf{w} = \mathbf{w} - \alpha \frac{\partial C}{\partial \mathbf{w}}$
  save relevant information
end

The gradient calculations should be done using matrix operations, not loops.