# Lecture 6: Logistic Regression
# CSC 84020 - Machine Learning

Andrew Rosenberg

February 19, 2009

# Last Time

- Regression
  - Regularization and Overfitting

# Today

- Logistic Regression

## Classification

Goal: Identify which of $K$ classes a data point $x$ belongs to.

Like Regression, Classification is a **supervised** task.

For each data point $\mathbf{x}_i$ we have a corresponding target (or label, or class) $t_i$ that describes the correct classification of the data point.

Goal: identify a function $y : \mathbb{R}^D \rightarrow C$ where
$t_i \in C = \{c_0, \ldots, c_{K-1}\}$

# Representations of the target variable

$$y : \mathbb{R}^D \rightarrow C \text{ where } t_i \in C$$

For binary (two-way) classification it is convenient to represent $t_i$ as a single scalar variable $t_i \in \{0, 1\}$.

- This will allow us to interpret $t_i$ as the likelihood that a point $\mathbf{x}_i$ is a member of class $c_{K-1}$
- When hypothesized from a model, this can represent the confidence of the prediction.

For K $\geq$ 2 classes, we represent $\mathbf{t}$ as a $K$ element vector, where, if a point is a member of class $c_j$ the $j$-th element is 1, and all the others are 0.

- In 5-way classification, a member of class $c_2$ is

$$\mathbf{t} = (0, 0, 1, 0, 0)^T$$

We may also represent $t$ as a nominal variable when using non-probabilistic models.

# Three approaches to Classification

**Generative Approach**

$$p(c_j|\mathbf{x}) = \frac{p(\mathbf{x}|c_j)p(c_j)}{p(\mathbf{x})}$$

**Discriminative approach**

$$p(c_j|\mathbf{x})$$

**Discriminant function**

$$f(\mathbf{x}) = c_j$$

# Three approaches to Classification

**Generative Approach** Highest resource requirements. Need to approximate the joint probability $p(\mathbf{x}, c_j)$

$$p(c_j|\mathbf{x}) = \frac{p(\mathbf{x}|c_j)p(c_j)}{p(\mathbf{x})}$$

**Discriminative approach**

$$p(c_j|\mathbf{x})$$

**Discriminant function**

$$f(\mathbf{x}) = c_j$$

# Three approaches to Classification

**Generative Approach**

$$p(c_j|\mathbf{x}) = \frac{p(\mathbf{x}|c_j)p(c_j)}{p(\mathbf{x})}$$

**Discriminative approach** Moderate resource requirements. Typically less parameters to approximate than **generative** models

$$p(c_j|\mathbf{x})$$

**Discriminant function**

$$f(\mathbf{x}) = c_j$$

# Three approaches to Classification

**Generative Approach**

$$p(c_j|\mathbf{x}) = \frac{p(\mathbf{x}|c_j)p(c_j)}{p(\mathbf{x})}$$

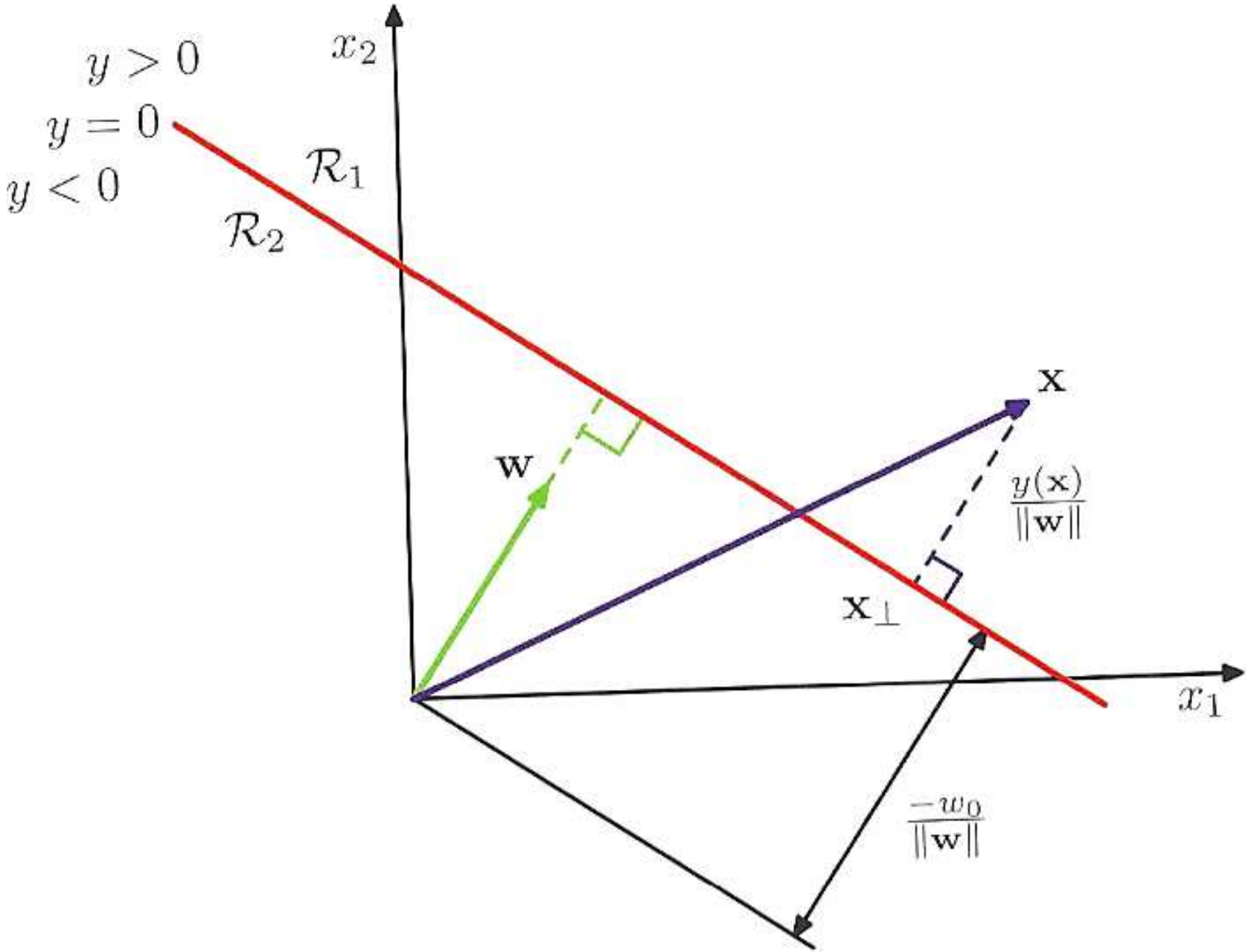
**Discriminative approach**

$$p(c_j|\mathbf{x})$$

**Discriminant function** Can be *trained* probabilistically, but the output does not include confidence information.

$$f(\mathbf{x}) = c_j$$

# Discriminant Functions



# Why Discriminant Functions are limiting

What can Generative and Discriminative approaches that Discriminant Functions cannot?
...Or why we like probabilities

- Minimizing Risk – continuous updating.
- Reject Option – "I don't know"
- Compensating for Priors
- Combining Models

We'll talk about these more when we discuss Perceptrons and Neural Networks.

# Generative Modeling

**Generative modeling** – model the posterior

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

# Generative Modeling

**Generative modeling** – model the posterior

$$
\begin{aligned}
p(c_1|\mathbf{x}) &= \frac{p(\mathbf{x}|c_1)p(c_1)}{p(x)} \\
&= \frac{p(\mathbf{x}|c_1)p(c_1)}{\sum_j p(x, c_j)}
\end{aligned}
$$

# Generative Modeling

**Generative modeling** – model the posterior

$$
\begin{aligned}
p(c_1|\mathbf{x}) &= \frac{p(\mathbf{x}|c_1)p(c_1)}{p(x)} \\
&= \frac{p(\mathbf{x}|c_1)p(c_1)}{\sum_j p(x, c_j)} \\
&= \frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}, c_0) + p(\mathbf{x}, c_1)}
\end{aligned}
$$

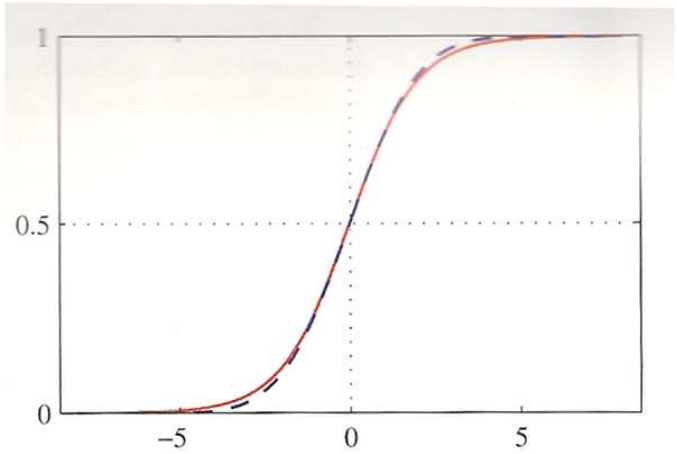
# Generative Modeling

**Generative modeling** – model the posterior

$$
\begin{aligned}
p(c_1|\mathbf{x}) &= \frac{p(\mathbf{x}|c_1)p(c_1)}{p(x)} \\
&= \frac{p(\mathbf{x}|c_1)p(c_1)}{\sum_j p(x, c_j)} \\
&= \frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}, c_0) + p(\mathbf{x}, c_1)} \\
&= \frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_0)p(c_0) + p(\mathbf{x}|c_1)p(c_1)}
\end{aligned}
$$

# Sigmoid function

The **sigmoid**[1] function is a "squashing function".




$$\sigma(x) = \frac{1}{1 + exp(-x)}$$

Squashing function maps the reals to a finite domain.

$$\sigma : \mathbb{R} \rightarrow (0, 1)$$

[1] "S-shaped"

# Generative Modeling

$$
\begin{aligned}
p(c_1|\mathbf{x}) &= \frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_0)p(c_0)+p(\mathbf{x}|c_1)p(c_1)} \\
&= \left(\frac{p(\mathbf{x}|c_0)p(c_0)+p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_1)p(c_1)}\right)^{-1} \\
&= \left(\frac{p(\mathbf{x}|c_0)p(c_0)}{p(\mathbf{x}|c_1)p(c_1)}+1\right)^{-1} \\
&= \left(\exp\left(\ln\frac{p(\mathbf{x}|c_0)p(c_0)}{p(\mathbf{x}|c_1)p(c_1)}\right)+1\right)^{-1} \\
&= \left(\exp\left(-\ln\frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_0)p(c_0)}\right)+1\right)^{-1} \\
a &= \ln\frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_0)p(c_0)} \\
p(c_1|\mathbf{x}) &= \frac{1}{1+\exp(-a)} \\
&= \sigma(a)
\end{aligned}
$$

# Some more vocabulary

*log-odds* or *log-odds-ratio*

$$a = \ln \frac{p(\mathbf{x}|c_1)p(c_1)}{p(\mathbf{x}|c_0)p(c_0)}$$

*logit* function – inverse of the sigmoid.

$$\begin{aligned} \sigma &= \frac{1}{1+\exp(-a)} \\ a &= \ln\left(\frac{\sigma}{1-\sigma}\right) \end{aligned}$$

# Generative Model

Derive $p(c_0|\mathbf{x})$ with Gaussian class conditional probability.

$$p(\mathbf{x}|c_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

We'll assume that $p(\mathbf{x}|c_0)$ and $p(\mathbf{x}|c_1)$ have equal covariance matrices.

$$\text{Want to show that } p(c_0|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$$

$$\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= \ln \frac{p(x|c_0)p(c_0)}{p(x|c_1)p(c_1)} \\
a &= \ln\left(\frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\}\right) \\
&\quad - \ln\left(\frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}\right) + \ln \frac{p(c_0)}{p(c_1)}
\end{aligned}$$

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= \ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\}\right) \\
&\quad -\ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\mathbf{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}\right)+\ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= \ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\}\right) \\
&\quad -\ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}\right)+\ln\frac{p(c_0)}{p(c_1)} \\
&= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= \ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right\}\right) \\
&\quad -\ln\left(\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}\right)+\ln\frac{p(c_0)}{p(c_1)} \\
&= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\ln\frac{p(c_0)}{p(c_1)} \\
&= -\frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0}-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0}-\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad +\frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1}-\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1}-\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad +\ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

If $\mathbf{A}$ is **symmetric** $\mathbf{A}=\mathbf{A}^\mathsf{T}$. If $\mathbf{A}$ is symmetric, $\mathbf{x}^T\mathbf{A}\mathbf{y}=\mathbf{y}^T\mathbf{A}\mathbf{x}$. (HW).

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= -\frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - \boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad + \frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad + \ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= -\frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - \boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad + \frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad + \ln\frac{p(c_0)}{p(c_1)} \\
a &= -\frac{1}{2}\left(\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - 2\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad + \frac{1}{2}\left(\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - 2\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad + \ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

If $\mathbf{A}$ is **symmetric** $\mathbf{A} = \mathbf{A^T}$. If $\mathbf{A}$ is symmetric, $\mathbf{x}^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{A}\mathbf{x}^T$. (HW).

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= -\frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - \boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad +\frac{1}{2}\left(\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad +\ln\frac{p(c_0)}{p(c_1)} \\
a &= -\frac{1}{2}\left(\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} - 2\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad +\frac{1}{2}\left(\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} - 2\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \\
&\quad +\ln\frac{p(c_0)}{p(c_1)} \\
a &= (\boldsymbol{\mu}_0\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1})\mathbf{x} - \frac{1}{2}\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} + \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

If $\mathbf{A}$ is **symmetric** $\mathbf{A} = \mathbf{A^T}$. If $\mathbf{A}$ is symmetric, $\mathbf{x}^T\mathbf{A}\mathbf{y} = \mathbf{y}^T\mathbf{A}\mathbf{x}^T$. (HW).

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= (\boldsymbol{\mu}_0\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1})\mathbf{x} - \frac{1}{2}\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} + \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

# Generative model

$$
\begin{aligned}
p(c_0|\mathbf{x}) &= \sigma(a) \\
a &= (\boldsymbol{\mu}_0\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_1\boldsymbol{\Sigma}^{-1})\mathbf{x} - \frac{1}{2}\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} + \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \ln\frac{p(c_0)}{p(c_1)} \\
a &= (\mathbf{w}^T\mathbf{x} + w_0) \\
\mathbf{w} &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0}^T - \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1}^T \\
w_0 &= -\frac{1}{2}\boldsymbol{\mu_0}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_0} + \frac{1}{2}\boldsymbol{\mu_1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_1} + \ln\frac{p(c_0)}{p(c_1)}
\end{aligned}
$$

# Maximum Likelihood Solution

Now we have a way to describe the linear transformation to $\mathbf{x}$ to generate a prediction under a Gaussian assumption.

How do we estimate the parameters?

Maximize the likelihood function with respect to each parameter.

$$p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = \prod_{n=0}^{N-1} (\pi N(\mathbf{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}))^{t_n} ((1-\pi) N(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{1-t_n}$$

- $t_n = 1$ for class 0, $t_n = 0$ for class 1.
- Prior class probabilities $p(C_0) = \pi$, $p(C_1) = 1 - \pi$.

# Maximum Likelihood Solution

Optimize $\pi$.

$$
\begin{aligned}
p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) &= \prod_{n=0}^{N-1} (\pi N(\mathbf{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}))^{t_n} ((1-\pi) N(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{1-t_n} \\
\ln p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) &= \ln \prod_{n=0}^{N-1} (\pi N(\mathbf{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}))^{t_n} ((1-\pi) N(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}))^{1-t_n} \\
&= \sum_{n=0}^{N-1} t_n \ln(\pi N(\mathbf{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})) + (1-t_n) \ln((1-\pi) N(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})) \\
&= \sum_{n=0}^{N-1} t_n \ln(\pi N(\mathbf{x}_n|\boldsymbol{\mu}_0, \boldsymbol{\Sigma})) + (1-t_n) \ln((1-\pi) N(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})) \\
&= \sum_{n=0}^{N-1} t_n \ln(\pi) + (1-t_n) \ln(1-\pi) + const
\end{aligned}
$$

# Maximum Likelihood Solution

$$
\begin{aligned}
\ln p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}) &= \sum_{n=0}^{N-1} t_n \ln(\pi) + (1 - t_n) \ln(1 - \pi) + const \\
\frac{\partial \ln p(\mathbf{t}, \mathbf{X}|\pi, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{\partial \pi} &= \frac{1}{\pi} \sum_{n=0}^{N-1} t_n - \frac{1}{1 - \pi} \sum_{n=0}^{N-1} (1 - t_n) = 0 \\
\frac{1}{\pi} \sum_{n=0}^{N-1} t_n &= \frac{1}{1 - \pi} \sum_{n=0}^{N-1} (1 - t_n) \\
\frac{1 - \pi}{\pi} \sum_{n=0}^{N-1} t_n &= \sum_{n=0}^{N-1} (1 - t_n) \\
\left( \frac{1}{\pi} - 1 \right) \sum_{n=0}^{N-1} t_n &= \sum_{n=0}^{N-1} (1 - t_n) \\
\left( \frac{1}{\pi} \right) \sum_{n=0}^{N-1} t_n &= \sum_{n=0}^{N-1} (1 - t_n) + \sum_{n=0}^{N-1} (t_n) \\
\left( \frac{1}{\pi} \right) \sum_{n=0}^{N-1} t_n &= \sum_{n=0}^{N-1} 1
\end{aligned}
$$

# Maximum Likelihood Solution

$$
\begin{aligned}
\left(\frac{1}{\pi}\right)\sum_{n=0}^{N-1} t_n &= \sum_{n=0}^{N-1} 1 \\
\left(\frac{1}{\pi}\right)\sum_{n=0}^{N-1} t_n &= N \\
\frac{1}{N}\sum_{n=0}^{N-1} t_n &= \pi \\
\frac{1}{N}\sum_{n=0}^{N-1} t_n &= \pi \\
\frac{N_0}{N} &= \pi \\
\frac{N_0}{N_0 + N_1} &= \pi
\end{aligned}
$$

Be prepared to maximize $\boldsymbol{\mu_0}$ and $\boldsymbol{\Sigma}$ for HW.

# Discriminative Linear Classification

In the generative case recall,

$$p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$$

- Can "generate" synthetic data from p(x).
- Need to model the joint probability

In Discriminative Modeling

- Model $p(t|\mathbf{x})$ directly

# Logistic Regression

From the generative case we can find that under some assumptions:

$$p(\mathbf{t}|\mathbf{x}) = y(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$$

In M-dimensions this has M parameters.
In the generative case 2M means and M(M+1)/2 covariance matrix[2].

Parameters grow linearly in M or quadratically in M.

So we'd rather optimize this function directly.

[2] Covariance matrices are symmetric

# Maximum likelihood

Define the Likelihood.

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=0}^{N-1} p(c_0|\mathbf{x}_n)^{t_n} \left\{1 - p(c_1|\mathbf{x}_n)^{1-t_n}\right\}$$

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=0}^{N-1} \{t_n \ln p(c_0|\mathbf{x}_n) + (1 - t_n) \ln p(c_1|\mathbf{x}_n)\}$$

Where $y_0 = p(c_0|\mathbf{x}_n) = \sigma(a_n)$.

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=0}^{N-1} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

This is also the *cross entropy* error function.[3]

[3] Logisitic Regression is also called **maximum entropy** or **maxent**.

# Maximum Likelihood

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=0}^{N-1} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

chain rule

$$\nabla_{\mathbf{w}} E = \sum_{n=0}^{N-1} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \nabla_{\mathbf{w}} a_n$$

# Maximum Likelihood

Derivation of $\frac{\partial E}{\partial y_n}$.

$$
\begin{aligned}
E(\mathbf{w}) &= -\sum_{n=0}^{N-1} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\} \\
\frac{\partial E}{\partial y_n}(\mathbf{w}) &= \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} \\
&= \frac{y_n(1-t_n) - t_n(1-y_n)}{y_n(1-y_n)} \\
&= \frac{y_n - y_n t_n - t_n + y_n t_n}{y_n(1-y_n)} \\
&= \frac{y_n - t_n}{y_n(1-y_n)}
\end{aligned}
$$

# Maximum Likelihood

Derivation of $\frac{\partial y_n}{\partial a_n}$.

$$
\begin{aligned}
\frac{d\sigma}{da} &= \frac{d\frac{1}{1+\exp(-a)}}{da} \\
&= \frac{d(1+\exp(-a))^{-1}}{da} \\
&= (-1)(1+\exp(-a))^{-2}(\exp(-a))(-1) \\
&= (1+\exp(-a))^{-2}(\exp(-a)) \\
&= \frac{1}{1+\exp(-a)}\left(\frac{\exp(-a)}{1+\exp(-a)}\right) \\
&= \frac{1}{1+\exp(-a)}\left(\frac{1+\exp(-a)-1}{1+\exp(-a)}\right) \\
&= \frac{1}{1+\exp(-a)}\left(\frac{1+\exp(-a)}{1+\exp(-a)} - \frac{1}{1+\exp(-a)}\right) \\
&= \sigma(1-\sigma)
\end{aligned}
$$

# Maximum Likelihood

Derivation of $\nabla_{\mathbf{w}} a_n$

$$a_n = \mathbf{w}^T \mathbf{x}$$

$$\nabla_{\mathbf{w}} a_n = \mathbf{x}_n$$

# Maximum Likelihood

Putting it all together

$$\begin{aligned}
\nabla_{\mathbf{w}} E &= \sum_{n=0}^{N-1} \frac{\partial E}{\partial y_n} \frac{\partial y_n}{\partial a_n} \nabla_{\mathbf{w}} a_n \\
&= \sum_{n=0}^{N-1} \frac{y_n - t_n}{y_n(1-y_n)} (y_n(1-y_n))\mathbf{x}_n \\
&= \sum_{n=0}^{N-1} (y_n - t_n)\mathbf{x}_n
\end{aligned}$$

Same as gradient of the sum of squares error in linear regression.

# How do we optimize this?

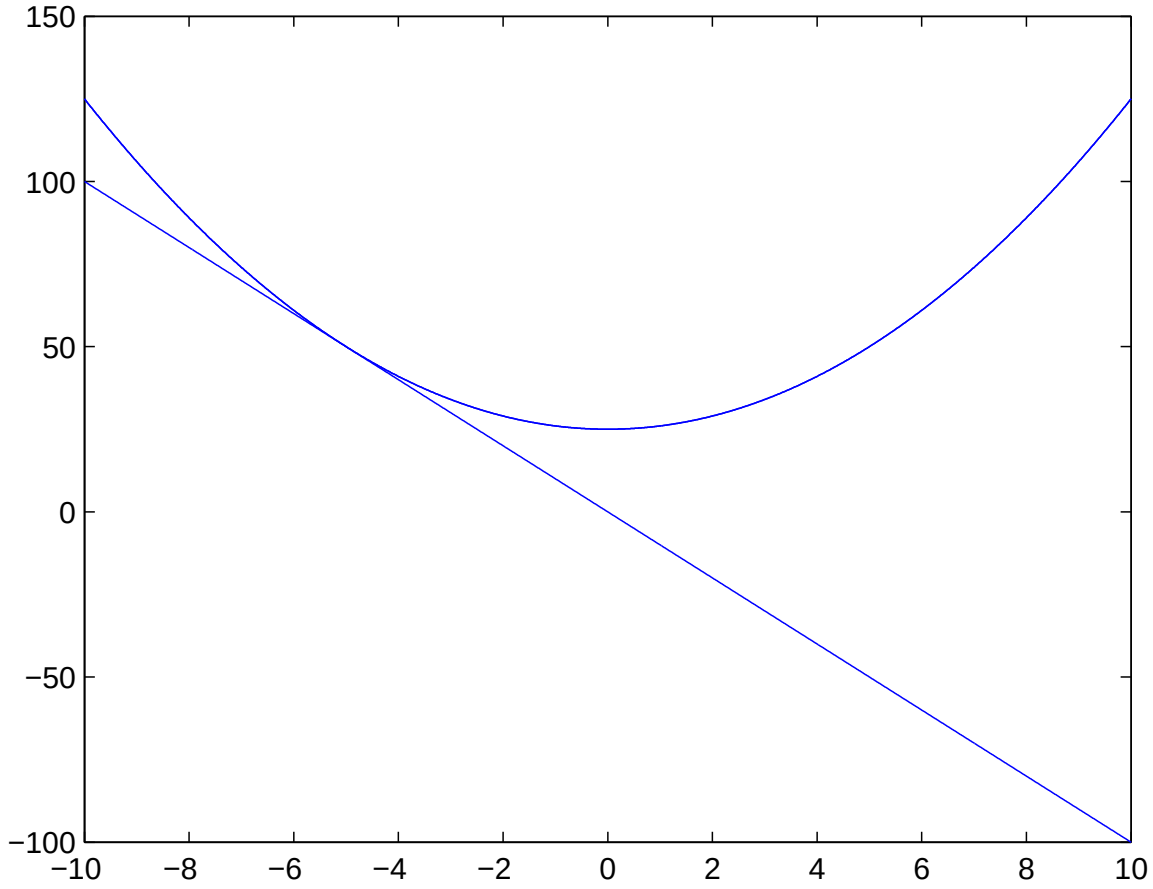
We know the gradient, but how do we find the maximum value

$$\nabla_{\mathbf{w}} E = \sum_{n=0}^{N-1} (y_n - t_n)\mathbf{x}_n$$

Numerical Approximations

- Gradient Ascent
  - $w_{n+1} = w_n + \eta \nabla_{\mathbf{w}} E(w_n)$
- Guess.
- Jump in the direction of the negative gradient.
- Guess again.

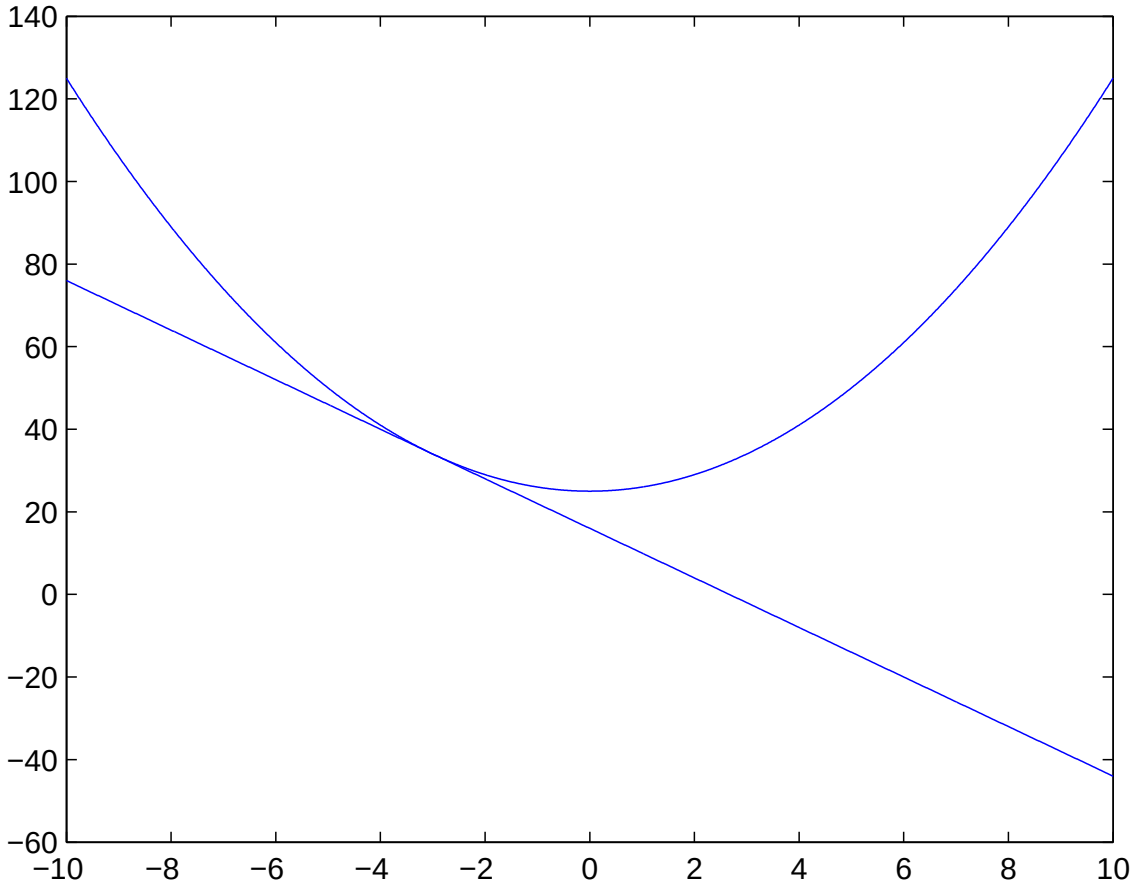
# Example of Gradient Descent




$x_0 = -5,\ f'(x_0) = -10,\ \eta = .2$

$x_1 = x_0 - \eta f'(x_0) = -5 - .2 * -10 = -3$

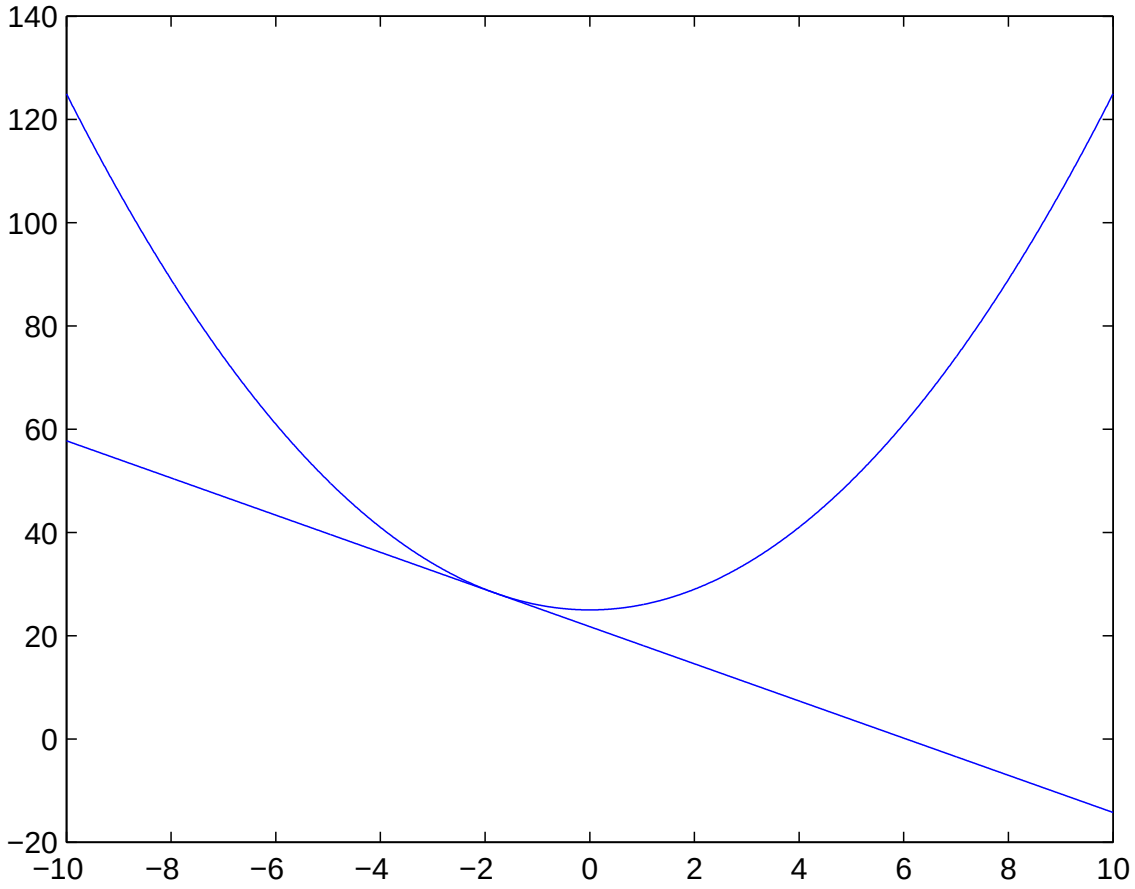
# Example of Gradient Descent




$x_1 = -3,\ f'(x_1) = -6,\ \eta = .2$

$x_2 = x_1 - \eta f'(x_1) = -3 - .2 * -6 = -1.8$

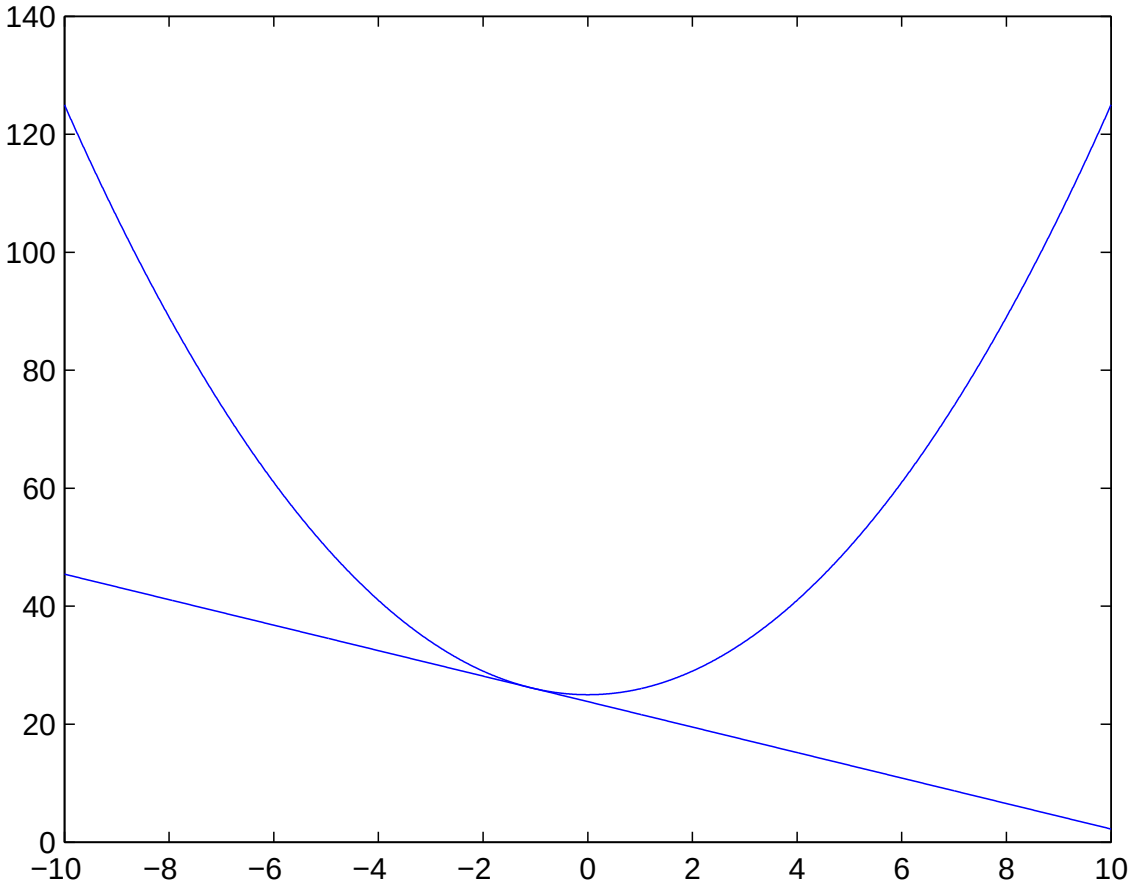
# Example of Gradient Descent




$x_2 = -1.8,\ f'(x_2) = -3.6,\ \eta = .2$

$x_3 = x_2 - \eta f'(x_2) = -1.8 - .2 * -3.6 = -1.08$

# Example of Gradient Descent




$x_3 = -1.08,\ f'(x_3) = -2.16,\ \eta = .2$

$x_4 = x_3 - \eta f'(x_3) = -5 - .2 * -10 = -.648$

# Another approach to N-way classification

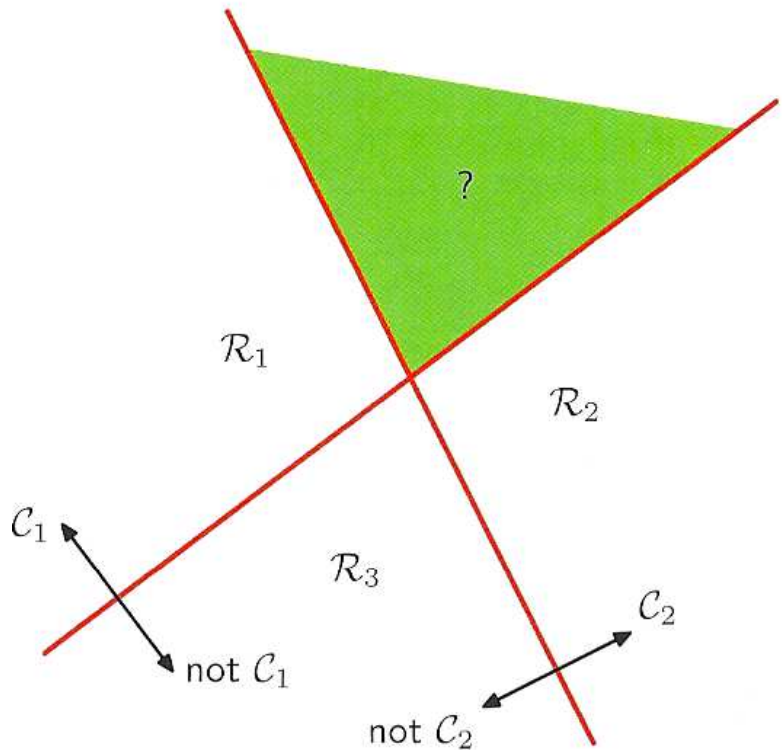
In this derivation we used a 1-of-$K$ representation with $K$-class discriminant function.

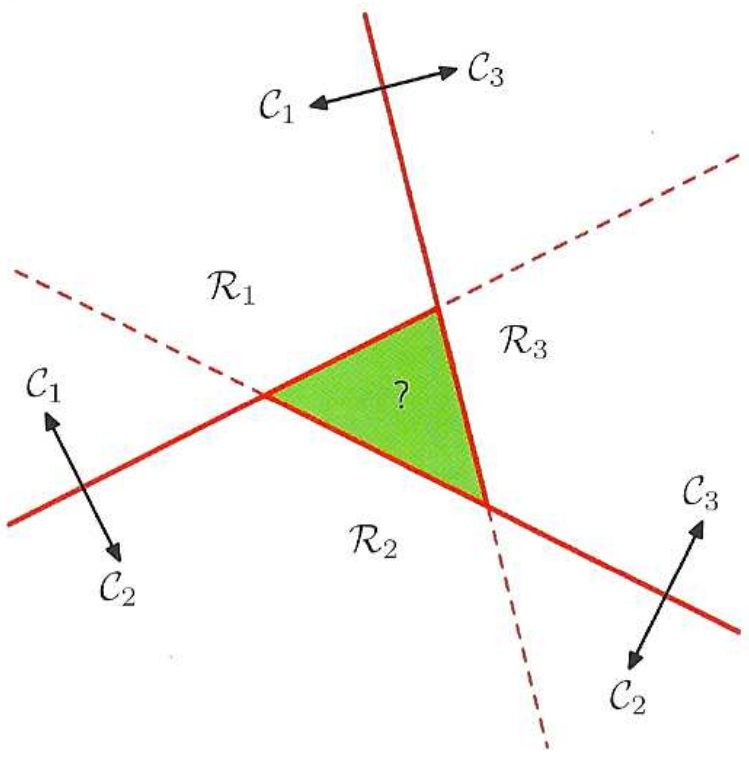
Another approach is to construct $K - 1$ binary classifiers.

- Each classifier $C_n$ compares $c_n$ to **not** $c_n$
- Binary Classifiers are simpler.

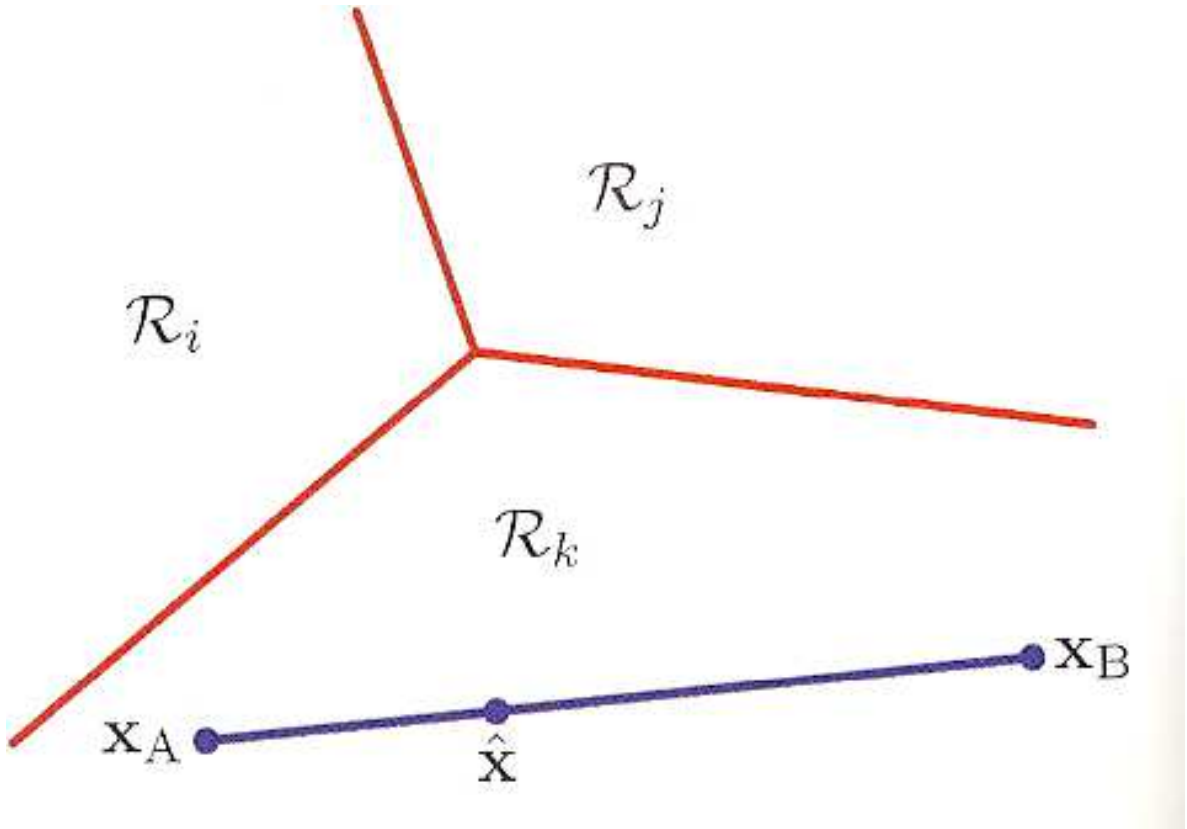
But there are some problems with this approach.

# One versus the rest



# $K$-class discriminant



# Context

Logistic Regression

- Powerful classification technique.
  - Must be approximated – no closed form.
- Assumption of linearity
- Can also be extended with *basis functions*.
- Also called **maximum entropy**.

# Bye

- Next
  - Graphical Models