Midterm Review (PART 2)

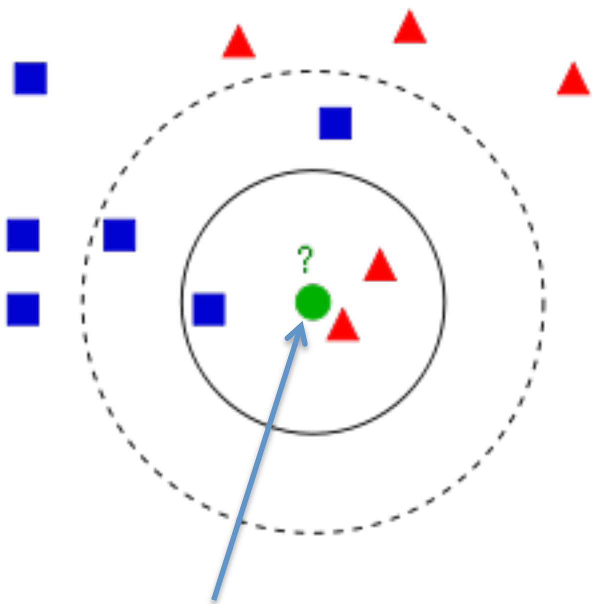
Topics Covered

1) K-NN (K Nearest Neighbor)
2) Linear Regression
3) Discriminant Analysis
4) Logistic Regression
5) Optimization / Convexity

K-Nearest Neighbor

What does it do?

Classifies objects based on the closest training example in feature space.



Test Sample

<u>Example</u>

If K=3, green circle will be assigned to the triangle class.

If K=5, green circle will be assigned to the square class.

<u>Naïve Metric utilizes Euclidean Distance</u>

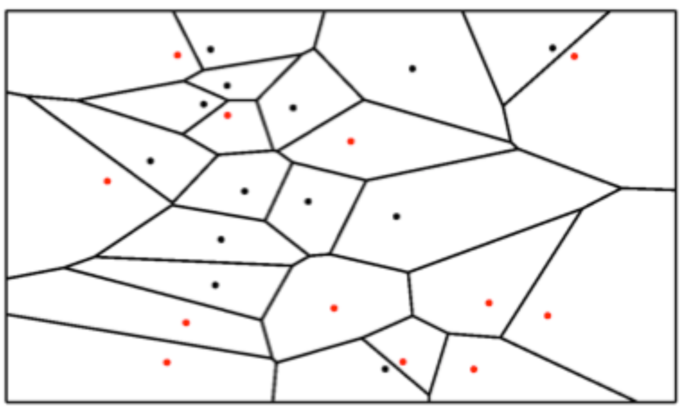
$$d(x_i, x_j) = ||x_i - x_j|| = \left(\sum_{\ell=1}^{784} (x_{i\ell} - x_{j\ell})^2 \right)^{0.5}$$

K-Nearest Neighbor

An example of a nonparametric model (parameters grow with the # of data points)

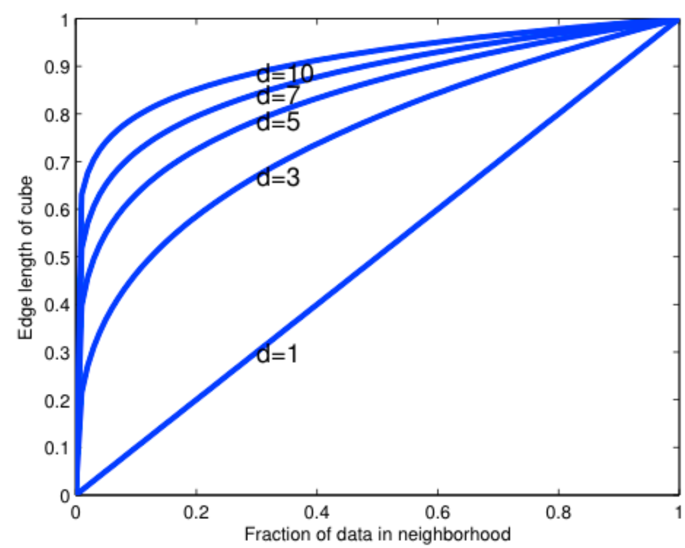
$$p(y=c|\mathbf{x},\mathcal{D},K) = \frac{1}{K}\sum_{i\in N_K(\mathbf{x},\mathcal{D})} \mathbb{I}(y_i = c)$$

Figure from Murphy (2012)



K=1 induces a Voronoi Tesselation

Curse of Dimensionality



Question: What happens to our K-NN classifier as the number of dimensions increases?

Linear Regression

We consider the problem of predicting real-valued outputs

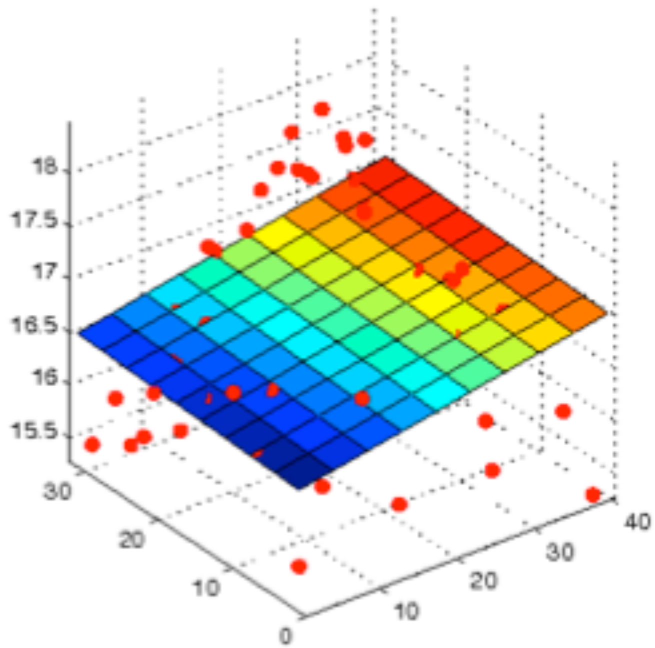


Figure from Murphy (2012)

Our prediction y is some linearly weighted combination of x plus some noise.

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

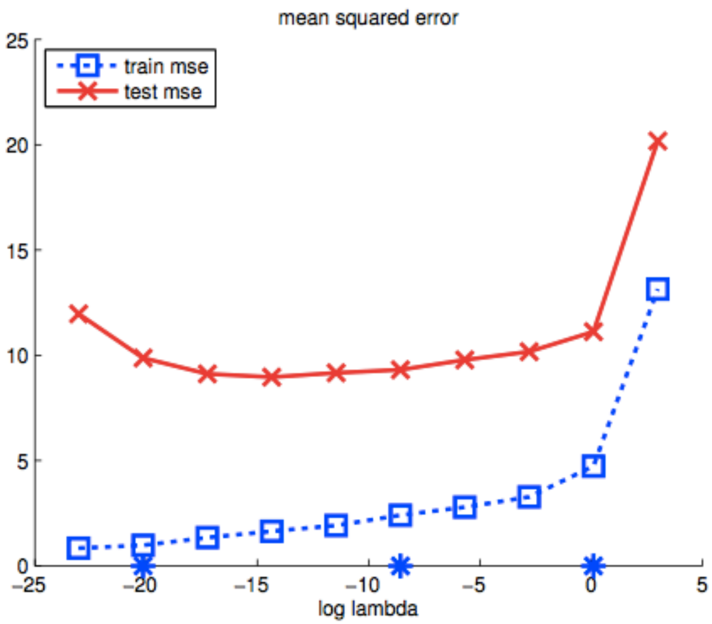
$$\mathbf{w}^T \mathbf{x} = \sum_{j=1}^{D} w_j x_j$$

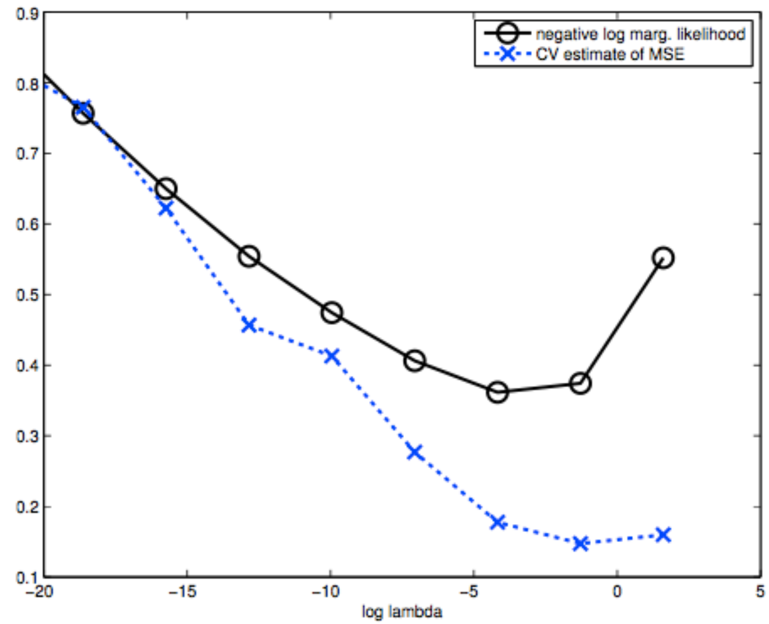
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Parametric or Nonparametric?
- What kind of a model (Generative or Discriminative?)

Linear Regression

Regularized Linear Regression (Ridge Regression)





Place a prior on <u>*w*</u>

$$p(\mathbf{w}) = \prod_j \mathcal{N}(w_j | 0, \tau^2)$$

Its inverse controls the strength of the prior.

MAP Estimation Problem

$$\underset{\mathbf{w}}{\operatorname{argmax}} \sum_{i=1}^{N} \log \mathcal{N}(y_i | w_0 + \mathbf{w}^T \mathbf{x}_i, \sigma^2) + \sum_{j=1}^{D} \log \mathcal{N}(w_j | 0, \tau^2)$$

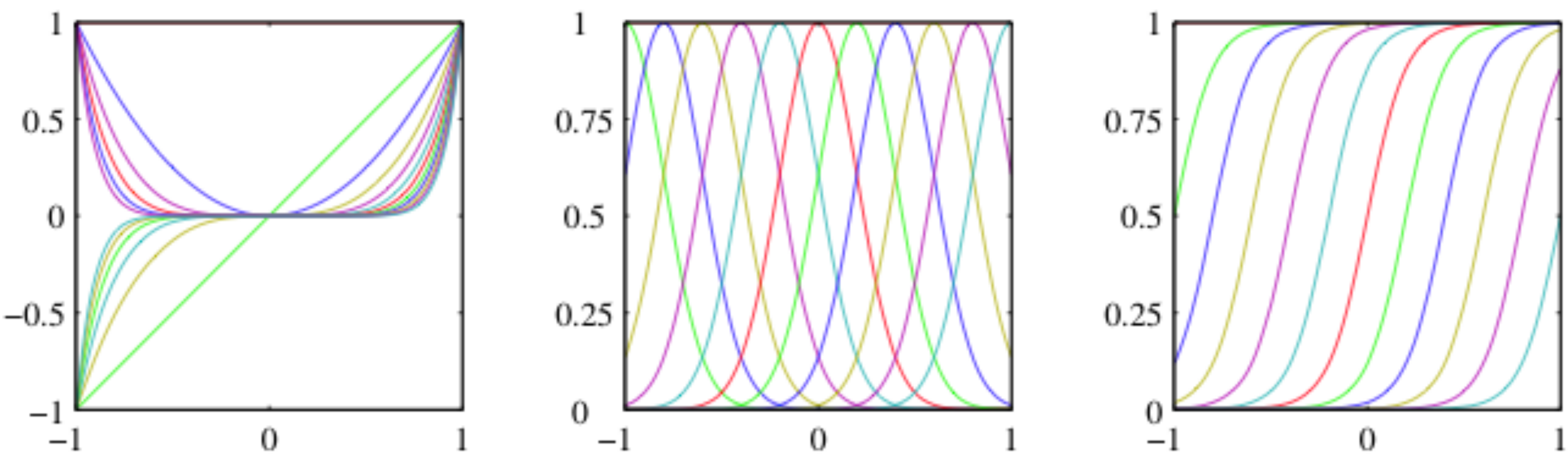
Linear Regression

Can also be written in this form, explicitly showing it as a conditional model

$$p(y \mid x, \theta) = N(y \mid w^T \phi(x), \sigma^2)$$

Basis Functions

Figure from Bishop (2006)



The MLE (Maximum Likelihood Estimate) has a unique solution

$$w_{ML} = (\phi^T \phi)^{-1} \phi^T y$$

Gaussian Discriminant Analysis

Our features are continuous and we wish to use a ***generative*** model for classification

Prior Term

$$p(y = c \mid \pi) = Cat(y = c \mid \pi)$$

Likelihood Term

$$p(x \mid y = c, \theta) = N(x \mid \mu_c, \Sigma_c)$$

Question: What is the MLE for these parameters?

Gaussian assumption on the class conditional densities

Deriving the posterior using Bayes Rule

$$p(y = c \mid \pi, x, \theta) = \frac{p(y=c|\pi)p(x|y=c,\theta)}{\sum_{c\prime=1}^{C} p(y=c\prime|\pi)p(x|y=c\prime,\theta)}$$

Quadratic Discriminant Analysis

The term ***quadratic discriminant analysis*** refers to the optimal decision boundaries for the posterior over our classes assuming a Gaussian conditional density for x.

$$p(y=c|\mathbf{x},\boldsymbol{\theta}) = \frac{\pi_c |2\pi\boldsymbol{\Sigma}_c|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right]}{\sum_{c'} \pi_{c'} |2\pi\boldsymbol{\Sigma}_{c'}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{c'})^T\boldsymbol{\Sigma}_{c'}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{c'})\right]}$$

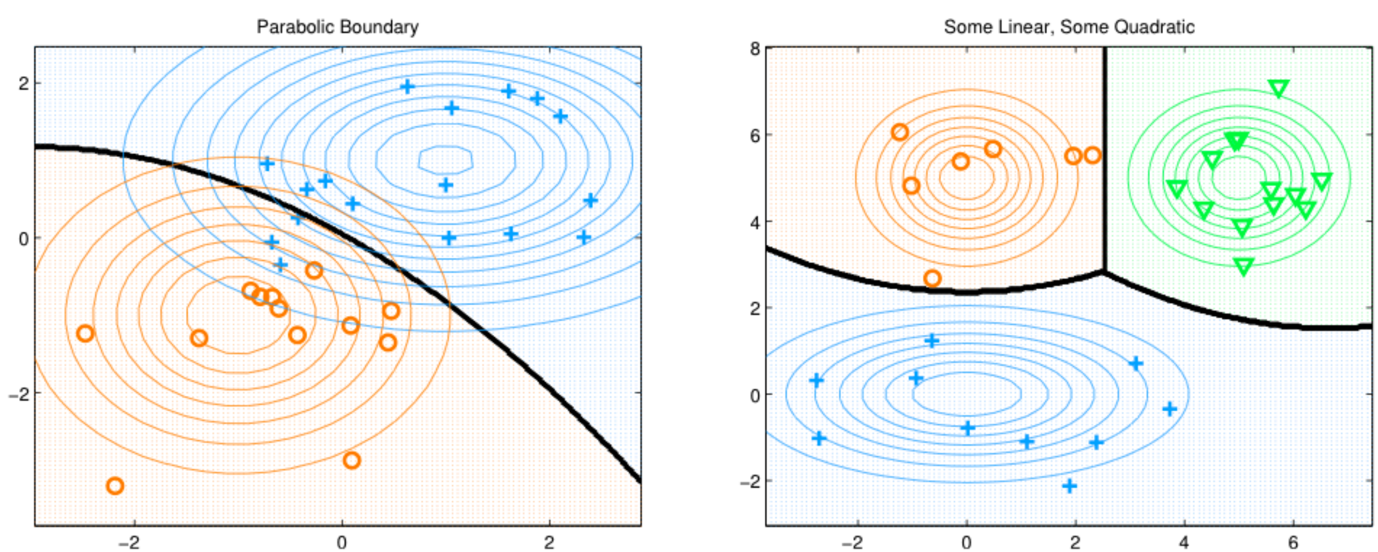


Figure from Murphy (2012)

Linear Discriminant Analysis

When our class conditional covariance parameters are all equal so that: $\Sigma_c = \Sigma$

The boundaries all become linear functions, thus the term ***Linear Discriminant Analysis***

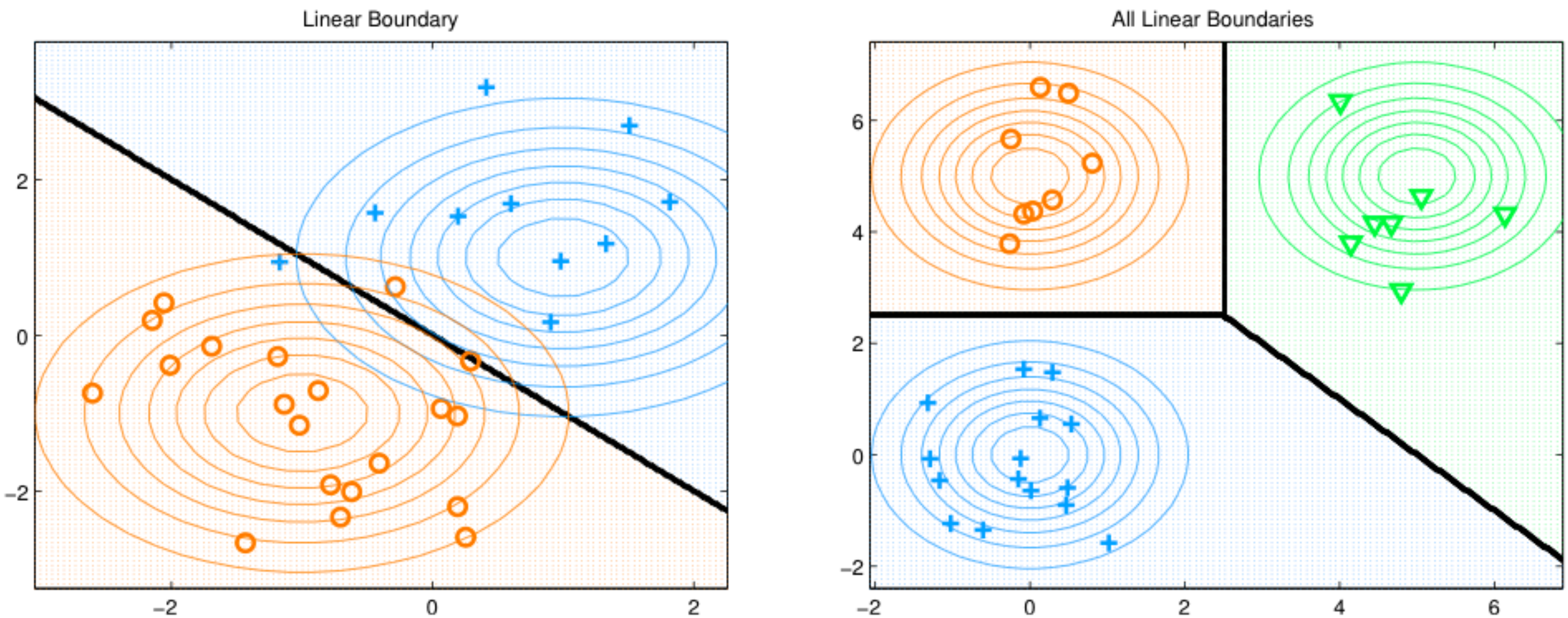
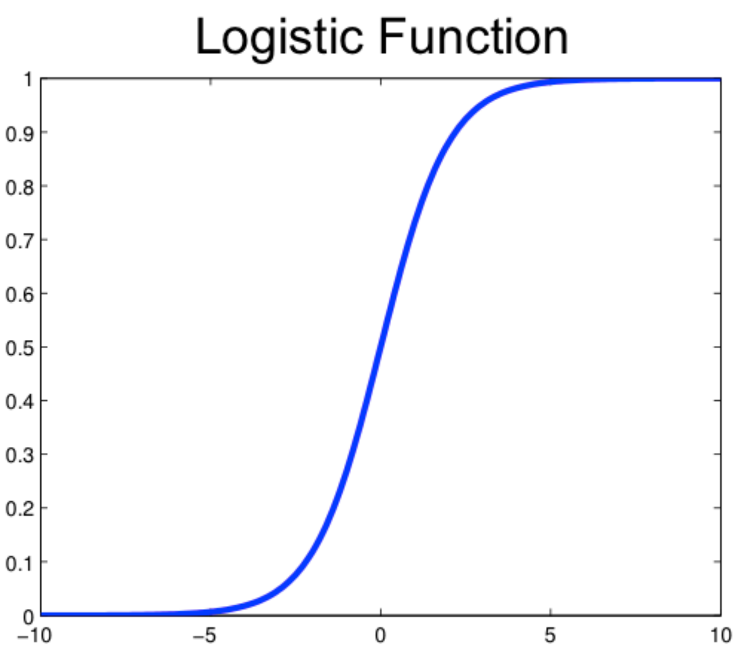


Figure from Murphy (2012)

Logistic Regression

Recall the Logistic Function

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



Returns a value between 0 and 1

Binary Logistic Regression

$$p(y \mid x, w) = \mathrm{Bern}(y \mid \sigma(w^T x))$$

Multi-Class Logistic Regression

$$p(y \mid x, W) = \mathrm{Cat}(y \mid S(W^T x))$$

Where S is the softmax function

$$S_c(W^T x) = \frac{\exp(w_c^T x)}{\sum_k \exp(w_k^T x)}$$

Question: Discriminative or Generative?

Question: Does this have a closed form solution?

Basis Functions in Linear Classification

Non-linear and Linear Boundaries from a Logistic Regression Model

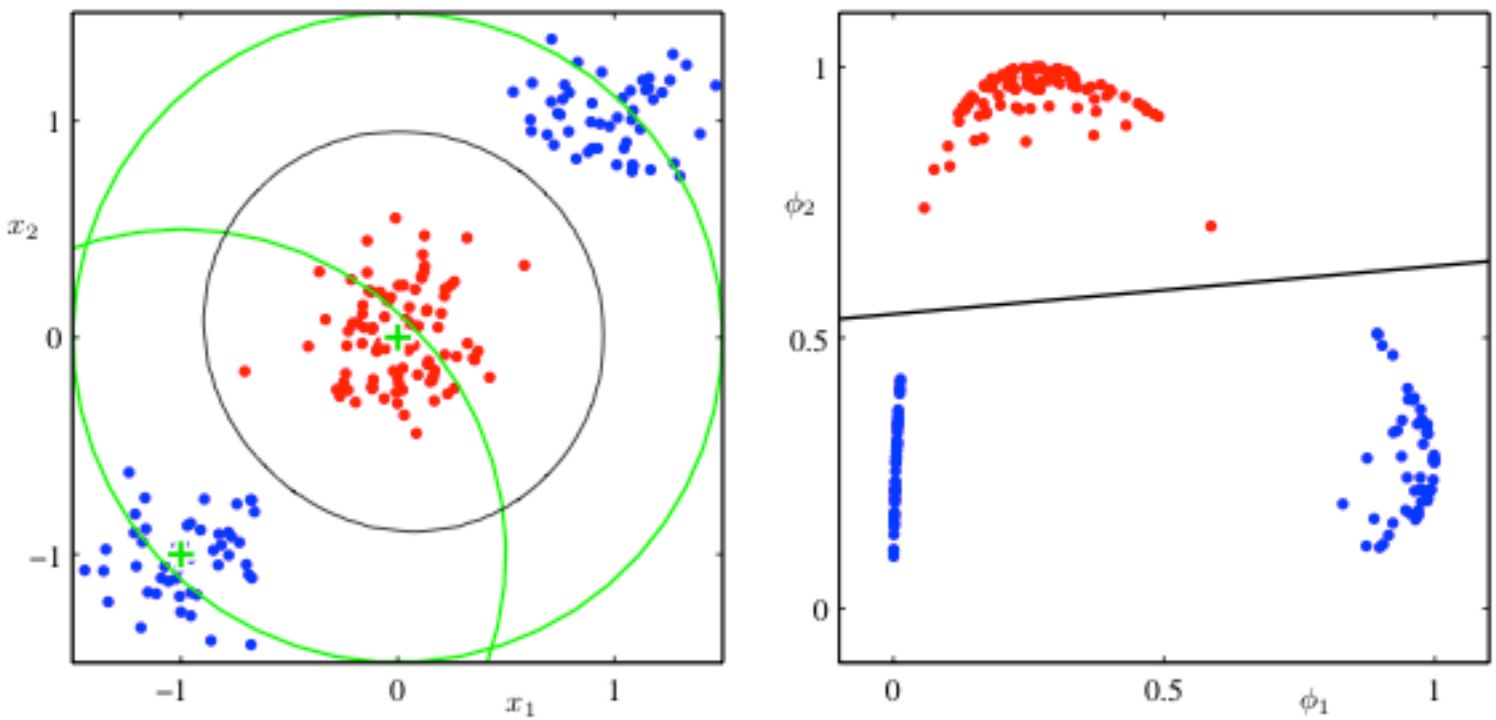


Figure from Bishop (2006)

Original input space for x

Input space mapped via
Gaussian basis functions

Convexity / Optimization

For Logistic Regression, we have no closed form solution for the optimal weights.

One Solution: Use optimization methods to find the best optimal value for w.

Basic Strategy for Logistic Regression:

1) Find the gradient / Hessian of our negative log likelihood.

$$\mu_i = \text{sigm}(\mathbf{w}^T \mathbf{x}_i)$$

$$\mathbf{g} = \frac{d}{d\mathbf{w}} f(\mathbf{w}) = \sum_i (\mu_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y})$$

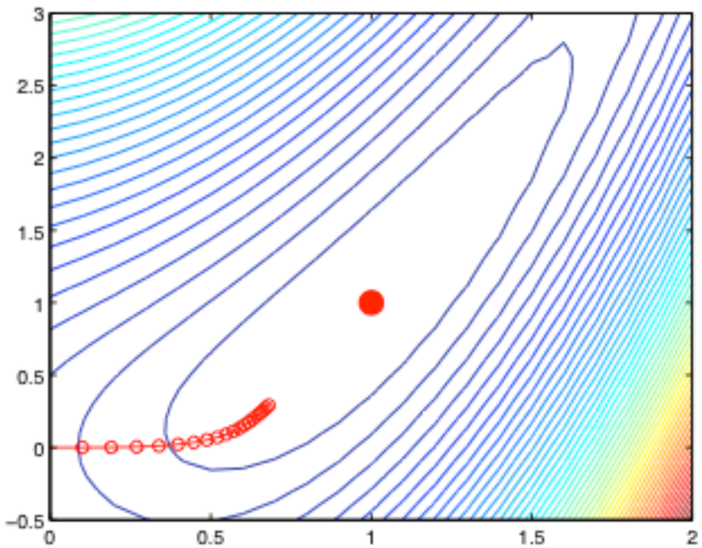
$$\mathbf{H} = \frac{d}{d\mathbf{w}} \mathbf{g}(\mathbf{w})^T = \sum_i (\nabla_{\mathbf{w}} \mu_i)\mathbf{x}_i^T = \sum_i \mu_i(1 - \mu_i)\mathbf{x}_i \mathbf{x}_i^T$$

2) Find optimal values for w, guided by information given to us by the gradient g and Hessian H.

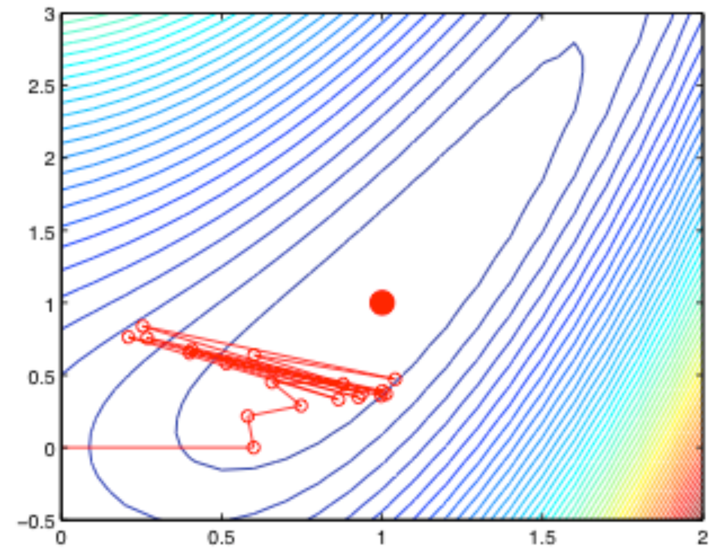
Optimization Techniques:

1) Steepest Descent (Uses only gradient information)
2) Newton's Method (Requires gradient and Hessian)
3) Quasi-Newton (Requires gradient only, estimates Hessian)

Steepest Descent



Stepsize = 0.1



Stepsize = 0.6

$$w_{k+1} = w_k - \eta_k \nabla_w f(w_k)$$

New value for "w" at iteration k + 1

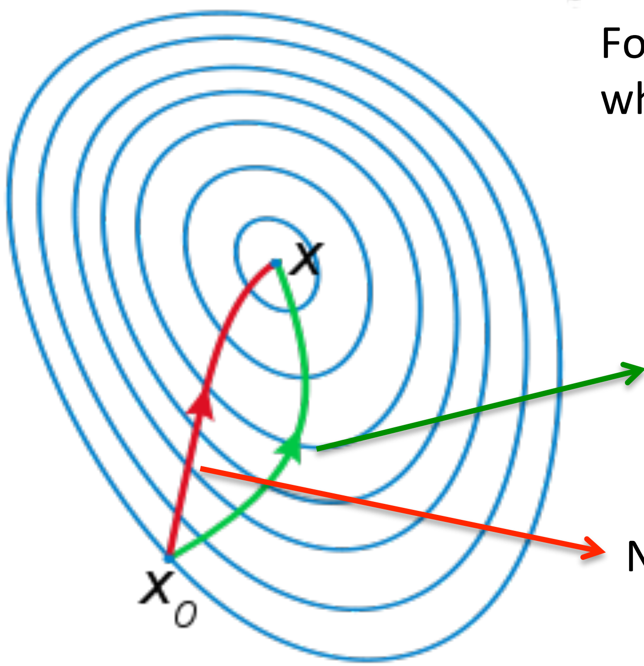
Stepsize

Gradient evaluated at w_k

Newton's Method

Uses second order information, the Hessian (curvature at f(x)) to find w.

For Logistic Regression, the Hessian is positive definite, which guarantees a convex optimization problem.



Gradient Method (slower)

Newton's Method (faster)

Notes:

1) Requires the inversion of the Hessian, which is computationally expensive.
2) For this method to work efficiently, Hessian must be positive definite.
3) When computing the inverse of H is too expensive, we can use Quasi-Newton methods.

Check out Ben's Demo for an interactive example!