Linear Classifiers

Agnes M Nielsen

DTU
Based on slides by Lars Arvastson and Line Clemmensen

02582 Computational Data Analysis, 2019

Today's Lecture

- Recap
- Linear Discriminant Analysis
- Logistic Regression
- Splines

Last week: Dimensionality

- Curse of Dimensionality
  - Number of regions grows exponentially with dimension
- Blessings of Dimensionality
  - Correlated features
  - Low-dimensional manifold
  - Underlying Structure

Last week: Regularization

- Methods
  - Ridge: $||\beta||_2$ shrinks parameters
  - Lasso: $||\beta||_1$ sets some parameters to zero
  - Elastic Net: Combines the other two
- Shrinkage and sparsity

Last week: Multiple Testing

- Testing many hypothesis
- Bonferroni Correction: Re-scale threshold with number of tests
- False Discovery Rate: Control the number of false discoveries

Linear Discriminant Analysis

Classification method

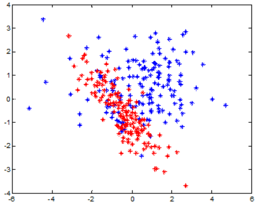
- Based on probability of class belonging
- Linear decision boundary

Linear Discriminant Analysis

- Classification from a probabilistic viewpoint
  - $P(G = k | X = x)$
  - Probability of class k, given observation x

Example:
$P(G = \text{red} | X = [0, -1])$ and
$P(G = \text{blue} | X = [0, -1])$

Predict that observation $X = [0, -1]$ belongs to the class with the **highest probability**



- We need a **stochastic model** for data to calculate probabilities
- Assume that data come from different **Gaussian distributions**
  - Different mean
  - Same correlation structure (just for simplicity)
- Data from different classes will overlap
- A straight line will be our decision boundary

Calculating Class Probabilities

$G(x)$ predicts class belonging for x,

$$G(x) = \arg\max_k \mathbf{P}(G = k | X = x).$$

Probablity given by **Bayes theorem**

$$\mathbf{P}(G = k | X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{k} f_\ell(x)\pi_\ell}$$

$f_\ell =$ distribution for class ℓ
$\pi_\ell =$ a priori probability for class ℓ (estimate or best guess)
Total probability, $\sum \pi_\ell = 1$.

Odds-Ratios

Look at log-**odds-ratio** for the two classes k and ℓ

$$\log \frac{\mathbf{P}(G=k|X=x)}{\mathbf{P}(G=\ell|X=x)} = \log \frac{f_k(x)\pi_k / \sum_i f_i \pi_i}{f_{\ell(x)}\pi_\ell / \sum_i f_i \pi_i}$$
$$= \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell}$$

We must make an assumption about f.

Assume that data in each class follows a multivariate normal distribution,

$$f(x) = (2\pi)^{-p/2} |\Sigma_k|^{-1/2} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-u_k)}$$

with a common covariance matrix $\Sigma_k = \Sigma$.

The Decision Boundary is Linear

$$
\begin{aligned}
\log & \frac{\mathbf{P}(G = k|X = x)}{\mathbf{P}(G = \ell|X = x)} \\
&= \ldots \\
&= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T \Sigma^{-1} (\mu_k - \mu_\ell) + x^T \Sigma^{-1} (\mu_k - \mu_\ell)
\end{aligned}
$$

Along the decision boundary we have $f_k \pi_k = f_\ell \pi_\ell$ (equal probability for both classes) and a log-odds-ratio $= \log 1 = 0$.

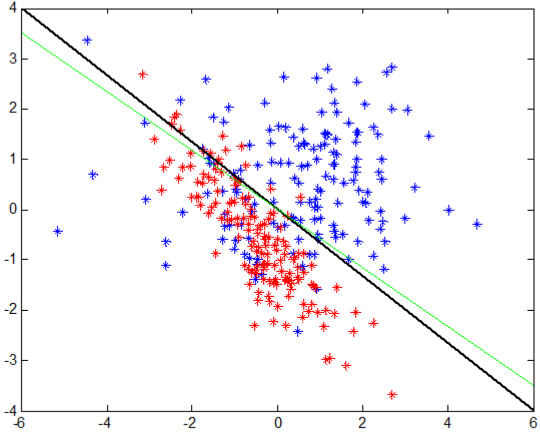
The decision boundary becomes

$$
\log \frac{\pi_k}{\pi_\ell} - \frac{1}{2}(\mu_k + \mu_\ell)^T \Sigma^{-1} (\mu_k - \mu_\ell) + x^T \Sigma^{-1} (\mu_k - \mu_\ell) = 0
$$

which is linear in x - in p dimensions a **hyper plane** like,

$$
a + x^T b = 0
$$

Example: Result From LDA



LDA in Practice

The decision rule $G(x)$ assigns class with highest probability

$$G(x) = \arg\max_k \delta_k(x).$$

using discriminant functions ($P(G = k|X = x)$ with constants removed)

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k; \qquad k = 1, ..., K$$

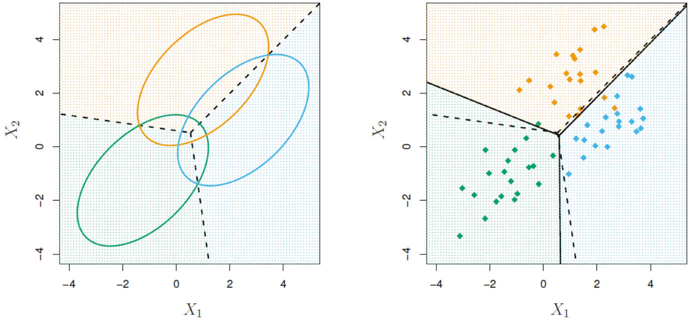
Use **plug-in estimates** for unknown parameters,

$$\hat{\pi}_k = N_k / N, \text{ where } N_k \text{ is number of class-k observations}$$

$$\hat{\mu}_k = \sum_{g_i = k} x_i / N_k$$

$$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K)$$

More Than Two Classes



- One decision line for each pair of classes
- One discriminant function for each class
  - Assign class to highest probability.

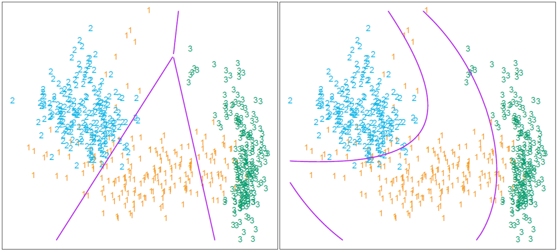
Variations of LDA

- Quadratic Discriminant Analysis
- Regularized Discriminant Analysis
- Reduced Rank Discriminant Analysis
- Sparse Discriminant Analysis

Quadratic Discriminant Analysis

LDA assumes that the covariance structures are equal.

When we drop this restriction we get **quadratic discriminant analysis**, QDA, and the decision boundaries becomes non-linear.



Regularized Discriminant Analysis

- In LDA when $p >> n$ the covariance matrix has rank at most n and needs to be inverted.
- The idea to solve this is similar to ridge regression. There are three ways.

Regularized Discriminant Analysis

It takes a lot of observations to estimate a large covariance matrix with precision. Three increasingly harsh regularizations are available

1. Make a compromise between LDA and QDA,

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

2. Shrink the covariance towards its diagonal

$$\hat{\Sigma}_k(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\,\mathrm{diag}(\hat{\Sigma})$$

3. Shrink the covariance towards a scalar covariance structure

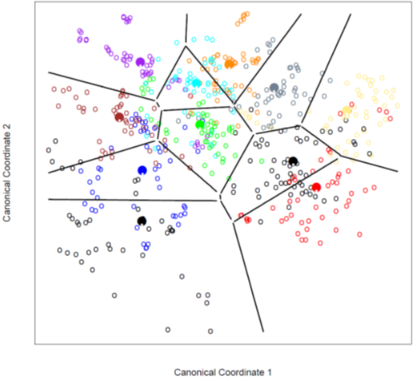
$$\hat{\Sigma}_k(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 I$$

Reduced Rank Discriminant Analysis

- The centroids lie in an affine subspace of dimension $\leq K - 1$
- Can we find a smaller space such that we can optimally view the LDA? i.e find the linear combination of the X such that the between-class variance is maximized relative to the within-class variance.

Reduced Rank Discriminant Analysis

Classification in a reduced subspace. In higher dimensional subspace, the decision boundaries are hyper-planes and can not be represented as lines. Hence, this technique is very **useful for illustrating class separation**.



Sparse Discriminant Analysis

It can be shown that optimal scoring and LDA are equivalent [ESL section 12.5]

$$\underset{\theta_k,\beta_k}{\text{minimize}} \, ||Y\theta_k - X\beta_k||_2^2$$

$$\text{s.t.} \quad \frac{1}{n}\theta_k^T Y^T Y\theta_k = 1, \quad \theta_k^T Y^T T\theta_l = 0 \quad \forall l < k$$

where θ_k is a K-vector of scores, β_k p-vector of parameters and Y is a matrix of dummy variables for the classes.

Sparse Discriminant Analysis

SDA comes from adding the ridge and lasso penalties

$$\underset{\theta_k,\beta_k}{\text{minimize}}\, ||Y\theta_k - X\beta_k||_2^2 + \lambda_2||\beta_k||_2^2 + \lambda_1||\beta_k||_1$$

$$\text{s.t.}\quad \frac{1}{n}\theta_k^T Y^T Y\theta_k = 1, \quad \theta_k^T Y^T T\theta_l = 0 \quad \forall l < k$$

This gives sparsity in features (good interpretation).

Logistic Regression

Classification method

- Fewer assumptions than LDA
- Linear decision boundary

Recall the LDA Assumptions

What made LDA linear?

- Equal covariance matrices
  - Unequal covariances lead to QDA
- Classes have Gaussian distributions

Logistic Regressions Throws Away These Assumptions

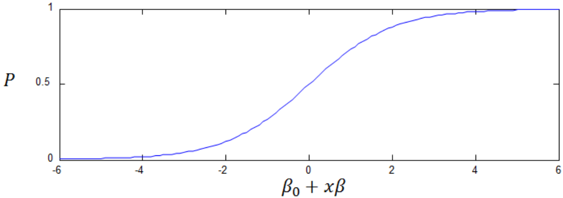
- Never mind about covariances and distributions!
- Optimize linear log-odds function directly
  - $\log \frac{P(G{=}red|X)}{P(G{=}blue|X)} = \beta_0 + x\beta$
- This is **logistic regression**
- What is a good choice of $\{\beta_0, \beta\}$?

Exercise: Two Class Problem

Derive expressions for the two class problem

$$P(G = \textcolor{red}{red} | X = x) = ?$$
$$P(G = \textcolor{blue}{blue} | X = x) = ?$$

when $\log \frac{P_r}{P_b} = \beta_0 + x\beta$



Likelihood Function

Combine this for all data points x_i

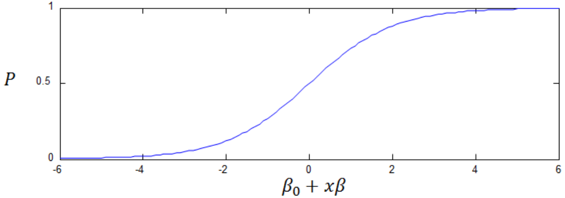
$$L(\beta_0, \beta) = \prod_{i=1}^{n} P(G = g_{x_i} | X = x_i)$$

Assuming independence, this is the **joint probability**

Fitting Logistic Regression

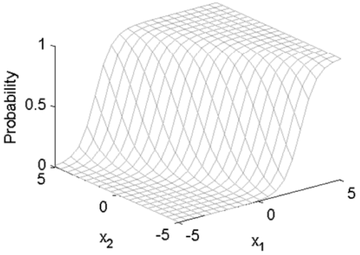
- Maximize the likelihood, L, wrt β_0 and β
  - $\arg\max_{\beta_0,\beta} L(\beta_0, \beta)$
  - Easier to maximize the log of $L(\beta_0, \beta)$
  - $l(\beta_0, \beta) = \log(L(\beta_0, \beta)) = \sum_i \mathbb{1}(x_i = \text{red})(\beta_0 + x_i\beta) - \log(1 + e^{\beta_0 + x_i\beta})$
- The approach is known as **maximum likelihood**
- The result is called **logistic regression**
- The maximization can be carried out using any method for numerical optimization
  - One algorithm uses an iteratively reweighted least squares solution

The Logistic Function

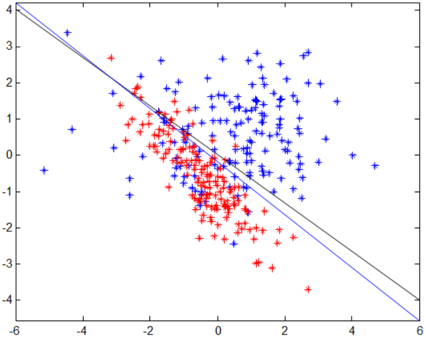


- $P(G = \textcolor{red}{red} | X = x) = \frac{e^{\beta_0 + x\beta}}{1 + e^{\beta_0 + x\beta}}$
- Decision boundary: $P = 1/2 \rightarrow \beta_0 + x\beta = 0$
- Well inside $P \approx 1$, well outside $P \approx 0$
  - Outliers are handled gracefully
  - Logistic regression focuses on observations close to the boundary

And In 2D



Logistic Regression vs. LDA



Multiple Logistic Regression

K classes, $K = 1, 2, ..., K$

$$\log \frac{P(G=1|X=x)}{P(G=K|X=x)} = \beta_{10} + x\beta_1$$

$$\log \frac{P(G=2|X=x)}{P(G=K|X=x)} = \beta_{20} + x\beta_2$$

$$\vdots$$

$$\log \frac{P(G=K-1|X=x)}{P(G=K|X=x)} = \beta_{(K-1)0} + x\beta_{K-1}$$

Arbitrary which class we put in the denominator

Multiple Logistic Regression

Since $P(G = K) = 1 - \sum_{i=1}^{K-1} P(G = i)$, we can show that

$$P(G = K|X = x) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + x\beta_i)}$$

and then

$$P(G = k|X = x) = \frac{\exp(\beta_{k0} + x\beta_k)}{1 + \sum_{i=1}^{K-1} \exp(\beta_{i0} + x\beta_i)}$$

Hence, the class probabilities does not depend on the choice of denominator in the odds-ratios.

Interpretation of Coefficients

- We have estimated β_0 and β
- What do they mean?
  - $\log \frac{P(G=red|X)}{P(G=blue|X)} = \beta_0 + x\beta$
  - They denote the log-odds contribution of each variable

Example: Model lung cancer (yes/no) as a function of smoking (number of cigarettes per day)

- $\beta = 0.02$
- A unit increase in smoking (one extra cigarette) means an increase in lung cancer risk (odds) of $\exp(0.02) \approx 1.02 = 2\%$

Regularized Logistic Regression

Few observations (low n) and high dimension (high p) data is a problem also for logistic regression.

One solution is an **elastic net** regularization of the likelihood,

$$
\begin{aligned}
[\beta, \beta_0] &= \arg\max_{\beta_0, \beta} \{\log L(\beta, \beta_0) - P_{\lambda,\alpha}(\beta)\} \\
&= \arg\max_{\beta_0, \beta} \left\{ \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{1+\beta_0+\beta^T x_i}) \right] - P_{\lambda,\alpha}(\beta) \right\}
\end{aligned}
$$

with

$$
P_{\lambda,\alpha}(\beta) = \lambda \left(\frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1 \right)
$$

Use cross-validation for λ and α.

Why Logistic Regression?

- Statistics
  - Identify variables important for separating the classes
  - For example in biostatistics and epidemiology
- Classification
  - Predict class belonging of new observations
  - For example spam/email or diseased/healthy
- Risk prediction
  - Estimate probability (risk) of each class
  - For example fraud detection in insurance claims

Some Properties of Logistic Regression

- Logistic regression is more robust than LDA
  - It relies on fewer assumptions
  - When is this a bad thing when compared to LDA?
- Logistic regression handles categorical variables better than LDA
- Observations far away from the boundary are down-weighted
- Breaks down when classes are perfectly separable
- Easy to interpret and explain
- Surprisingly often hard to beat
- Can be combined with regularization of parameters ($n < p$)
- Can be generalized to multi-class problems

Supervised Classification Overview

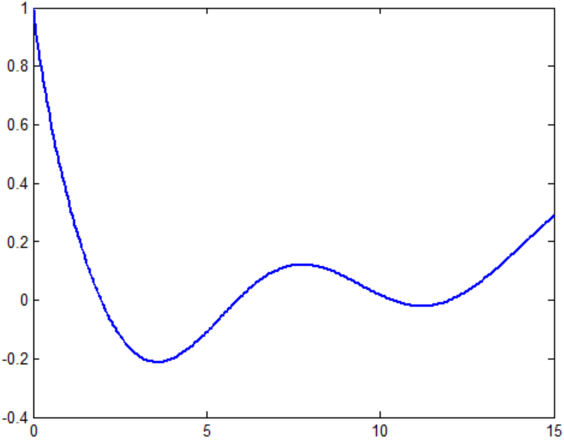
Overview of classification methods - so far.

Method	$p > n$	Dimension Reduction	Variable Selection	Weights on Observations	Assuming Independence	Distributional Assumptions	Equal Covariances	Non-linear Decision Boundary
Linear DA (LDA)						x	x	
Reduced-rank LDA		x				x	x	
Quadratic DA (QDA)						x		x
Logistic Regression				x				
K-Nearest Neighbor (KNN)				x				x
Regularized DA	x	x	x		(x)	x	x	
Sparse DA	x	x	x			x	x	

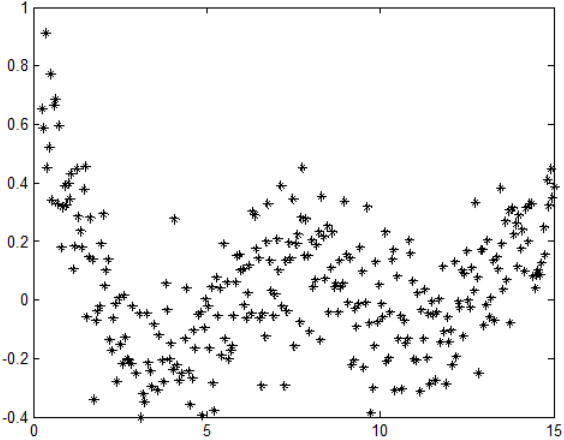
Basis Expansions and Splines

- Non-linear transformations
- Cubic splines

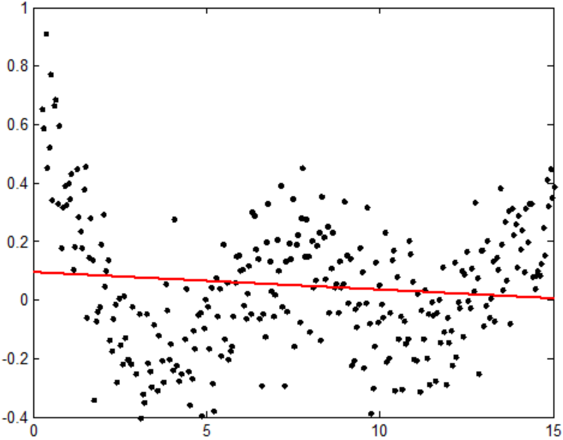
For a Second Return to Regression..



Observed Data Would look Something Like This



Ordinary Least Squares



Transform to Be Able to Use Linear Model

Idea: replace variables (columns) of the data matrix, X, with transformations $h(X)$

The linear model

$$y = X\beta = \sum_{i=1}^{p} \beta_i x_i \rightarrow \sum_{i=1}^{M} \beta_i' h_i(X)$$

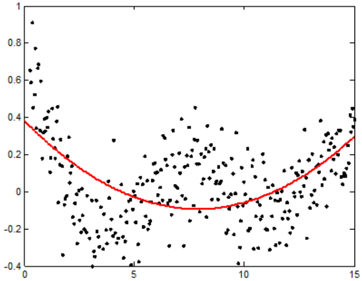
In this way we handle non-linear problems with our well known linear models

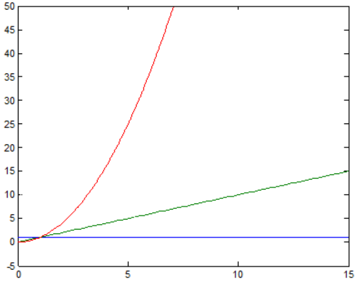
Basis Expansion

- We are not limited to use our data as they are
- Linear models
  - Easy to interpret
  - First order Taylor expansion of non-linearities
  - Might be ok even for non-linear data if we have few observations
- Non-linear problem - transform data and use linear model
  - $h_m(X) = X_j^2$ and $h_m(X) = X_j X_k$
  - $h_m(X) = \log(X_j)$ or $h_m(X) = sqrt(X_j)$
  - $h_m(X) = \frac{X - m_X}{s_X}$ (always used when using regularization)
  - $h_m(X_{(i)}) = i$, sort data $X_{(1)} \leq X_{(2)} \leq ...$ and use the rank
  - Either replacing X with $h_m(X)$ or expanding $\{X, h_m(X)\}$

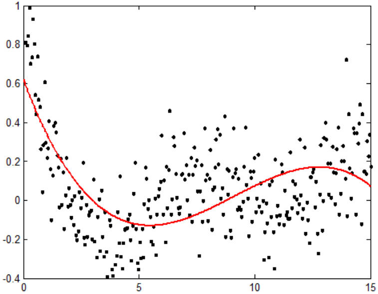
Example: Let's Try With Up to X^2

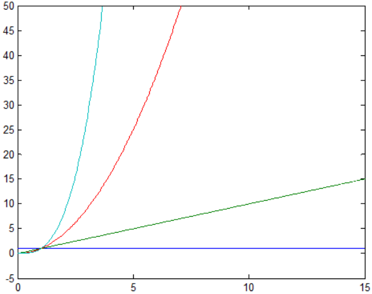
```
X = [ones(n,1) x x.^2];
y = X*(X\y)
```



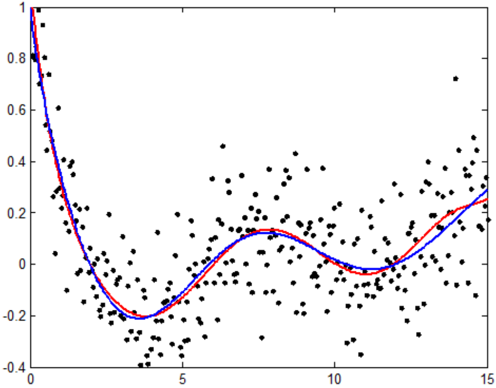


Example: Or With Up to X^3





Example: Or With Up to X^8

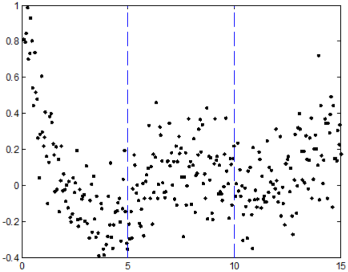


More advanced Transformations

- Splines
  - We'll talk about that next
- Fourier/Wavelet transforms
  - Time series data/images
- Principal components
  - Projection along eigenvectors
- Moving averages
  - Possibly also delayed averages capturing time dynamics
  - Time series data

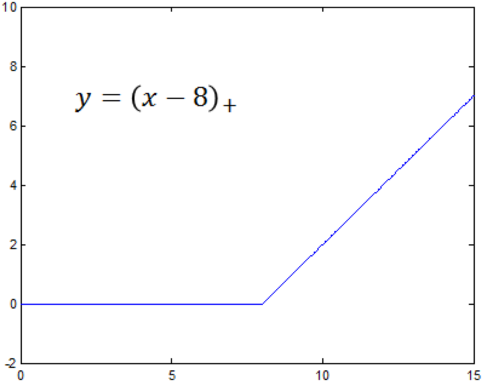
Piece-Wise

- To introduce flexibility while keeping the variance under control, we define **different basis functions for different intervals of x.**
- Example, divide the range of x in three parts



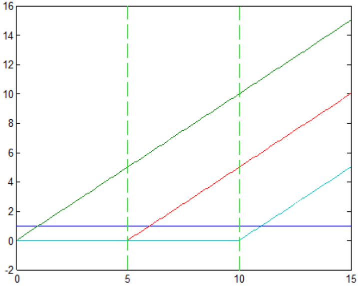
This Can Be Achieved Using The Hinge Function

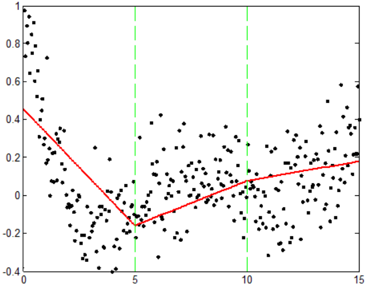
- Introducing the "**hinge**" function $y = (f(x))_+$
- Zero when $f(x)$ is less than zeros, otherwise $f(x)$



Example: Try with Piece-Wise Linear Functions

```
X = [ones(n,1) x max(0,x-5) max(0,x-10)];
```





Cubic Splines

Splines are piece-wise polynomials

- The basis functions are

```
X = [ones(n,1) x x.^2 x.^3 max(0,x-5).^3 max(0,x-10).^3];
```

- Or in the proper way

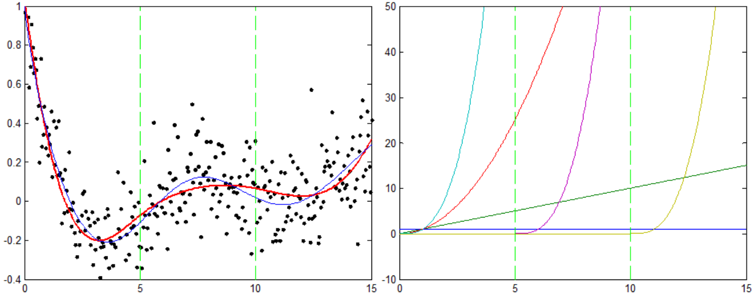
$$h_0(x) = 1$$
$$h_1(x) = x$$
$$h_2(x) = x^2$$
$$h_3(x) = x^3$$
$$h_4(x) = (x-5)_+^3$$
$$h_5(x) = (x-10)_+^3$$

- Cubic splines have continuous first and second derivatives at the knots.
  - I.e a smooth function

Example: Try With Cubic Splines

Spline approximation is

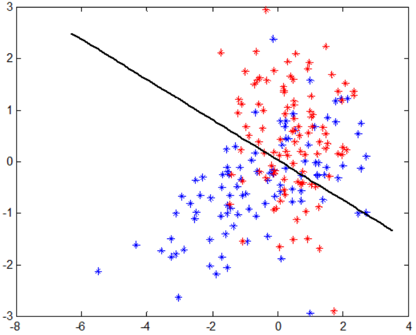
$$f(t) = b_0 + b_1 t + b_2 t^2 + b_3 t^3 + b_4(t-5)_+^3 + b_5(t-10)_+^3$$



Notice that this non-linear function was obtained with a linear model

Return to Linear Discriminant Analysis

Recall this gives linear decision boundary but we can apply our new knowledge!

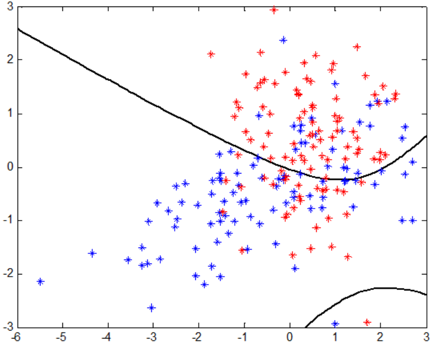


LDA with Basis Expansion

The decision boundary became non-linear in x, despite a linear classifier

Contour plot, $x\beta^T + \beta_0 = 0$

```
Xe = [ X X(:,1).^2 X(:,1).*X(:,2) X(:,2).^2 ];
```



Can Also Be Applied to Logistic Regression

The linear log-odds model is replaced with a flexible spline function

$$\log \frac{P(G=0|X=x)}{P(G=1|X=x)} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x-2)_+^3 + \beta_5 (x-5)_+^3$$

- Non-linear in x, linear in β
  - Standard logistic regression problem after basis expansion
- Easy to interpret
- Gives probability for class belongings

Summary: Linear Classifiers

- Linear Discriminant Analysis (LDA)
  - Linear decision boundary
  - Relies on Gaussian and equal covariances assumptions
  - Non-equal gave QDA and non-linear boundary
  - Regularized versions
- Logistic Regression
  - Linear decision boundary
  - Fewer assumptions than LDA
  - Coefficients can be interpreted
  - Regularized versions
- Basis Expansions
  - Gives linear methods non-linear decision boundaries
  - Cubic splines gave flexibility with variance under control