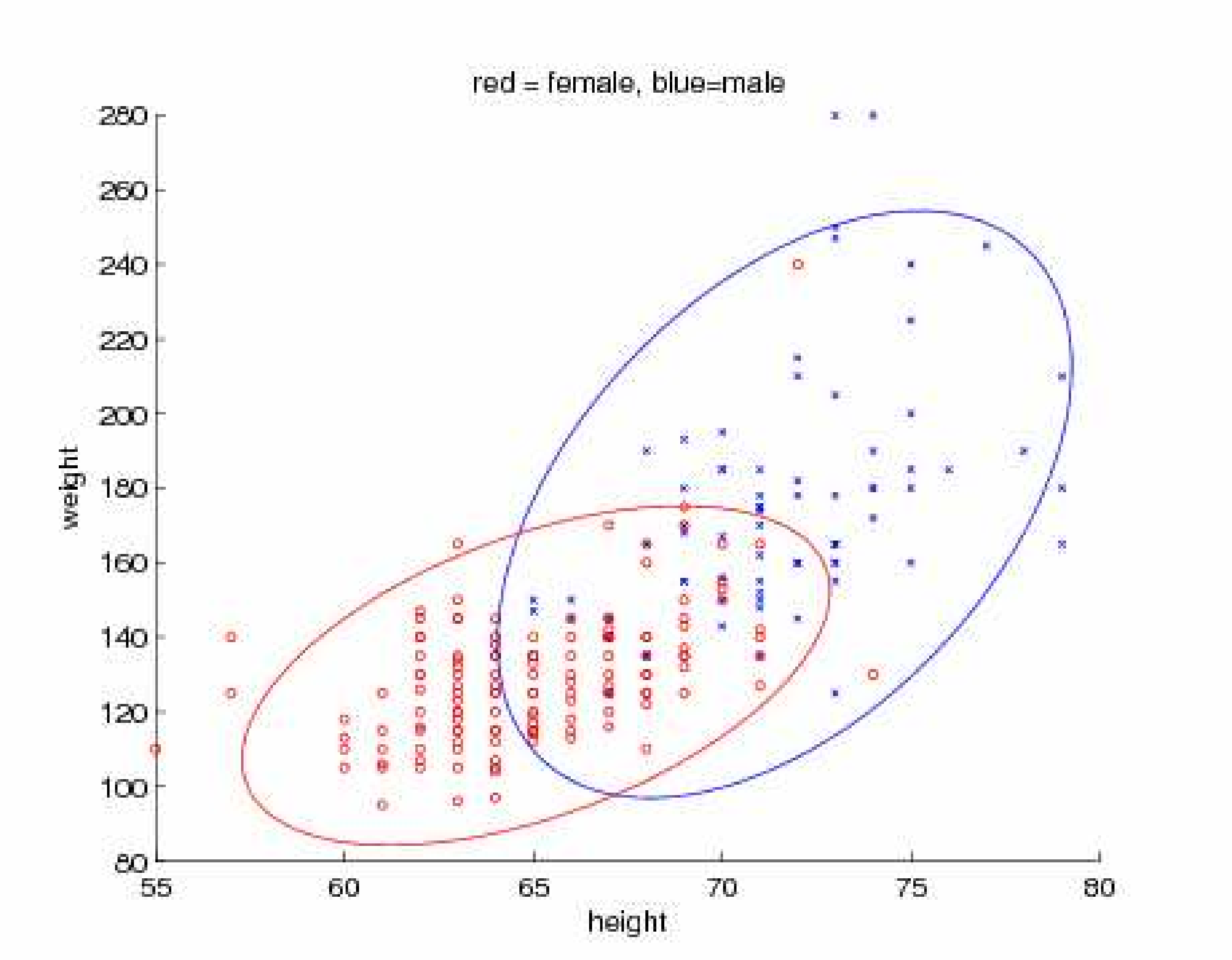
CS340 Machine learning
Gaussian classifiers

Correlated features

- Height and weight are not independent



Multivariate Gaussian

- Multivariate Normal (MVN)

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

- Exponent is the Mahalanobis distance between x and μ

$$\Delta = (\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Σ is the covariance matrix (positive definite)

$$\mathbf{x}^T\Sigma\mathbf{x} > 0 \;\forall \mathbf{x}$$

Bivariate Gaussian

- Covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

where the correlation coefficient is

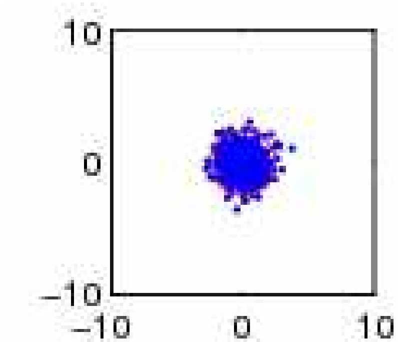
$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

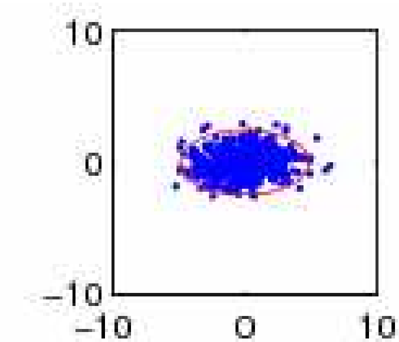
and satisfies -1 $\leq$ ρ $\leq$ 1

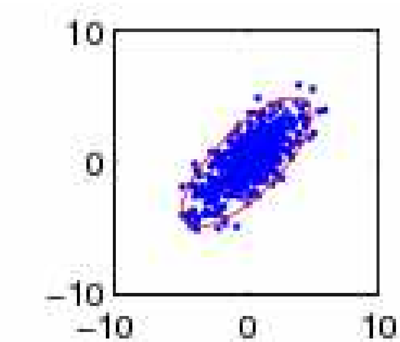
- Density is

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{(\sigma_x\sigma_y)}\right)\right)$$

Spherical, diagonal, full covariance

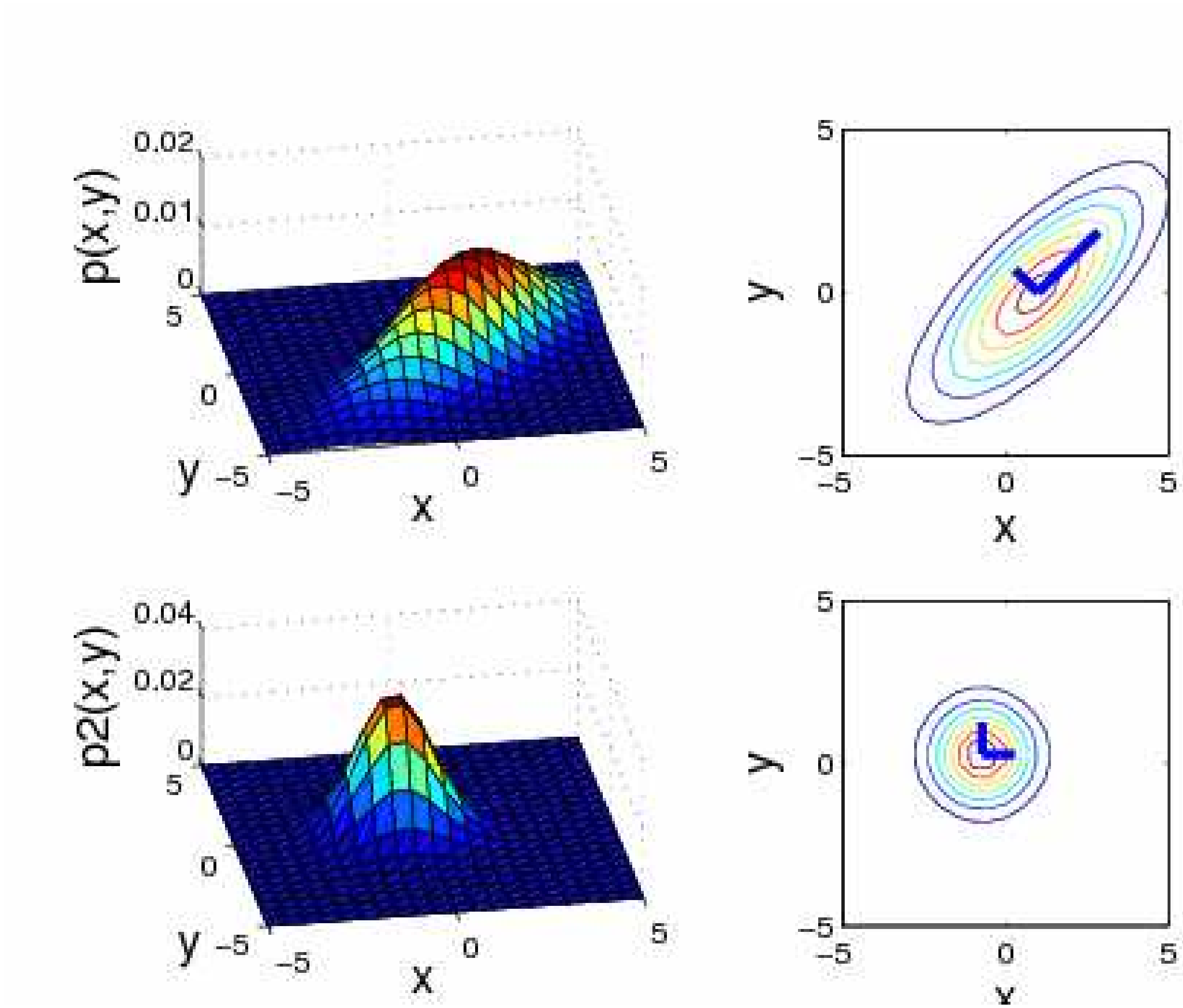






$$\Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

Surface plots



Generative classifier

- A generative classifier is one that defines a class-conditional density p(x|y=c) and combines this with a class prior p(c) to compute the class posterior

$$p(y=c|\mathbf{x}) = \frac{p(\mathbf{x}|y=c)p(y=c)}{\sum_{c'} p(\mathbf{x}|y=c')p(c')}$$

- Examples:
  - Naïve Bayes: $$p(\mathbf{x}|y=c) = \prod_{j=1}^{d} p(x_j|y=c)$$
  - Gaussian classifiers $$p(\mathbf{x}|y=c) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$$
- Alternative is a discriminative classifier, that estimates p(y=c|x) directly.

Naïve Bayes with Bernoulli features

- Consider this class-conditional density

$$p(x|y=c) = \prod_{i=1}^{d} \theta_{ic}^{I(x_i=1)} (1-\theta_{ic})^{I(x_i=0)}$$

- The resulting class posterior (using plugin rule) has the form

$$p(y=c|x) = \frac{p(y=c)p(x|y=c)}{p(x)} = \frac{\pi_c \prod_{i=1}^{d} \theta_{ic}^{I(x_i=1)} (1-\theta_{ic})^{I(x_i=0)}}{p(x)}$$

- This can be rewritten as

$$\begin{aligned} p(Y=c|x,\theta,\pi) &= \frac{p(x|y=c)p(y=c)}{\sum_{c'} p(x|y=c')p(y=c')} \\ &= \frac{\exp[\log p(x|y=c) + \log p(y=c)]}{\sum_{c'} \exp[\log p(x|y=c') + \log p(y=c')]} \\ &= \frac{\exp\left[\log \pi_c + \sum_i I(x_i=1)\log\theta_{ic} + I(x_i=0)\log(1-\theta_{ic})\right]}{\sum_{c'} \exp\left[\log \pi_{c'} + \sum_i I(x_i=1)\log\theta_{i,c'} + I(x_i=0)\log(1-\theta_{ic})\right]} \end{aligned}$$

Form of the class posterior

- From previous slide

$$p(Y = c|x, \theta, \pi) \quad \propto \quad \exp\left[\log \pi_c + \sum_i I(x_i = 1) \log \theta_{ic} + I(x_i = 0) \log(1 - \theta_{ic})\right]$$

- Define

$$\begin{aligned} x' &= [1, I(x_1 = 1), I(x_1 = 0), \ldots, I(x_d = 1), I(x_d = 0)] \\ \beta_c &= [\log \pi_c, \log \theta_{1c}, \log(1 - \theta_{1c}), \ldots, \log \theta_{dc}, \log(1 - \theta_{dc})] \end{aligned}$$

- Then the posterior is given by the softmax function

$$p(Y = c|x, \beta) = \frac{\exp[\beta_c^T x']}{\sum_{c'} \exp[\beta_{c'}^T x']}$$

- This is called softmax because it acts like the max function when $|\beta_c| \to \infty$

$$p(Y = c|\mathbf{x}) = \begin{cases} 1.0 & \text{if } c = \arg\max_{c'} \beta_{c'}^T \mathbf{x} \\ 0.0 & \text{otherwise} \end{cases}$$

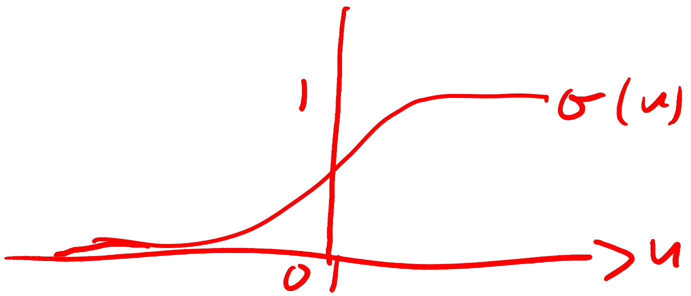
Two-class case

- From previous slide

$$p(Y=c|x,\beta) = \frac{\exp[\beta_c^T x']}{\sum_{c'} \exp[\beta_{c'}^T x']}$$

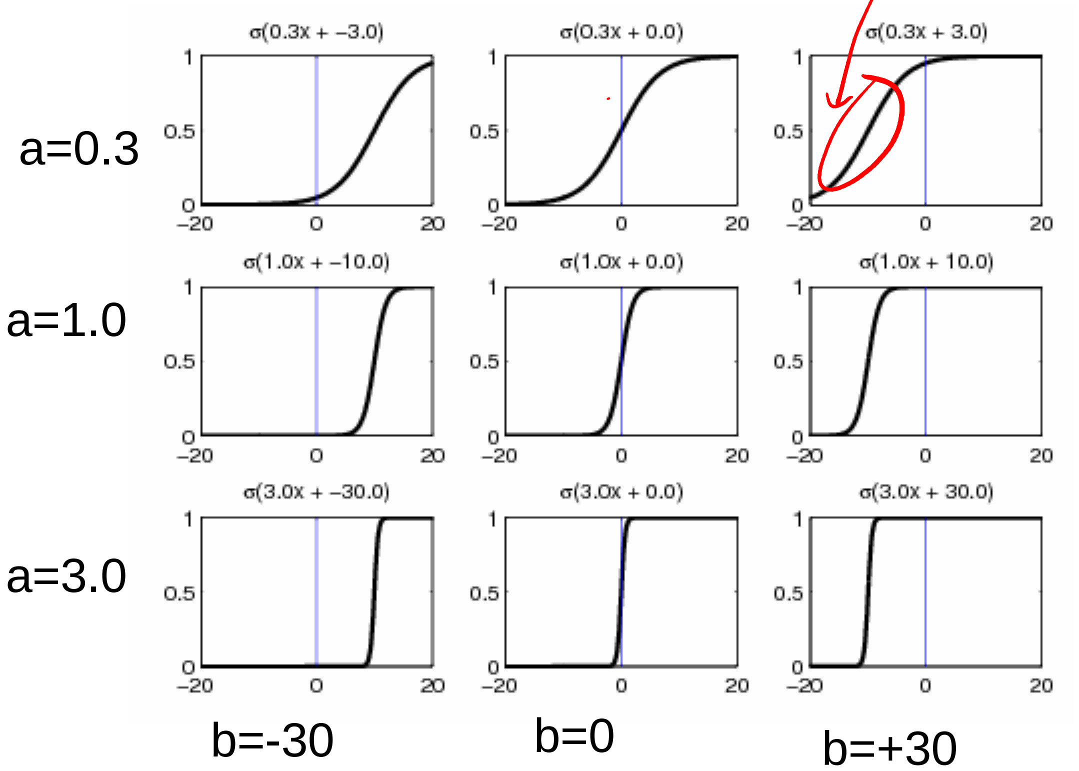
- In the binary case, Y $\in$ {0,1}, the softmax becomes the logistic (sigmoid) function $\sigma(u) = 1/(1+e^{-u})$

$$\begin{aligned} p(Y=1|x,\theta) &= \frac{e^{\beta_1^T x'}}{e^{\beta_1^T x'} + e^{\beta_0^T x'}} \\ &= \frac{1}{1+e^{(\beta_0-\beta_1)^T x'}} \\ &= \frac{1}{1+e^{w^T x'}} \\ &= \sigma(w^T x') \end{aligned}$$



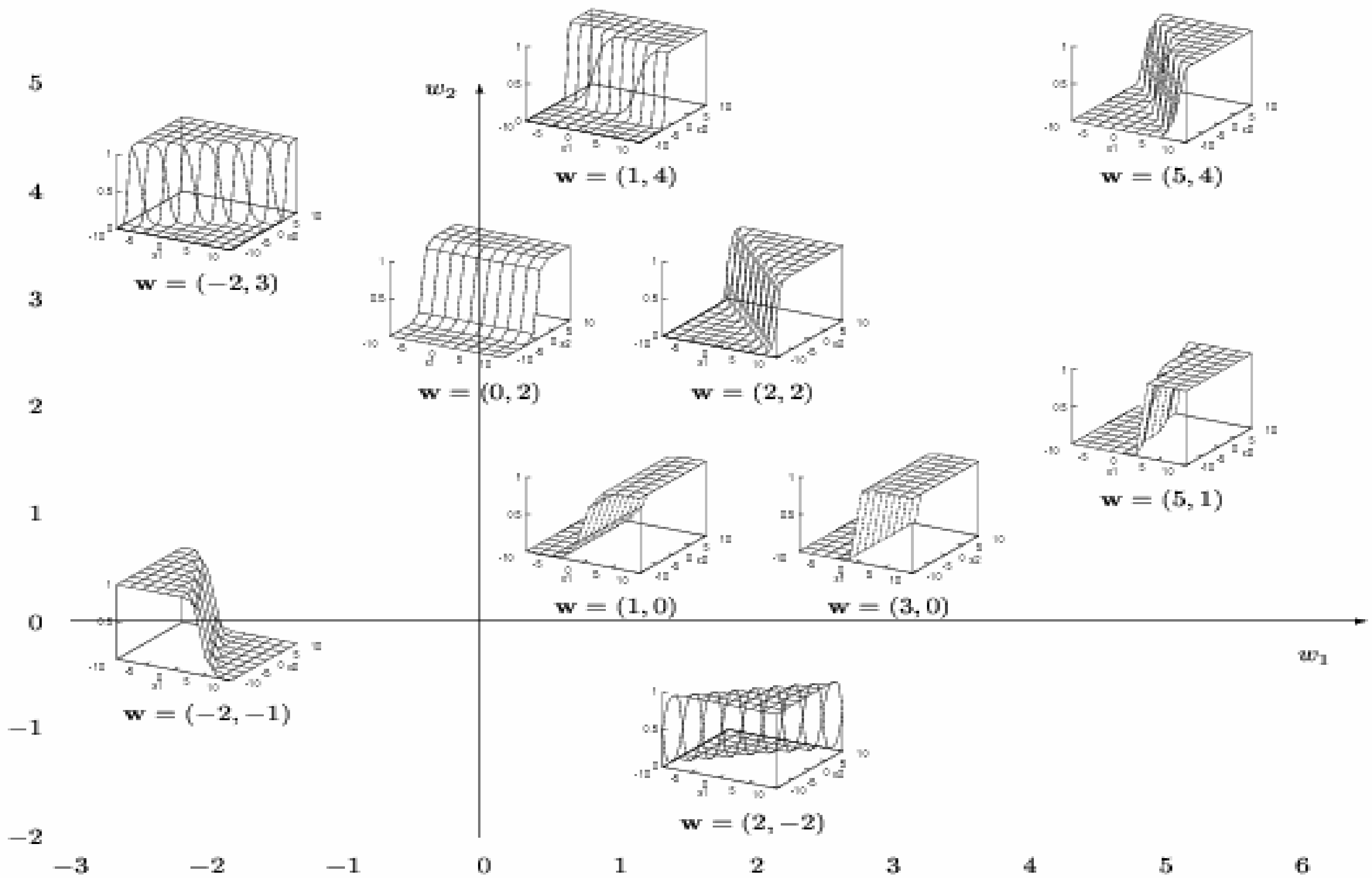
Sigmoid function

- $\sigma(ax + b)$, a controls steepness, b is threshold.
- For small a and $x \approx -b/2$, roughly linear



Sigmoid function in 2D

$\sigma(w_1\, x_1 + w_2\, x_2) = \sigma(w^T\, x)$: w is perpendicular to the decision boundary



Mackay 39.3

Logit function

- Let p=p(y=1) and η be the log odds

$$\eta = \log \frac{p}{1-p}$$

- Then p = σ(η) and η = logit(p)

$$\begin{aligned} \sigma(\eta) &= \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{e^{\eta}+1} \\ &= \frac{\frac{p}{(1-p)}}{\frac{p}{1-p}+1} = \frac{\frac{p}{(1-p)}}{\frac{p+1-p}{1-p}} = p \end{aligned}$$

η is the *natural parameter* of the Bernoulli distribution, and p = E[y] is the *moment parameter*

- If $\eta = w^T x$, then w_i is how much the log-odds increases by if we increase x_i

Gaussian classifiers

- Class posterior (using plug-in rule)

$$
\begin{aligned}
p(Y=c|\mathbf{x}) &= \frac{p(\mathbf{x}|Y=c)p(Y=c)}{\sum_{c'=1}^{C} p(\mathbf{x}|Y=c')p(Y=c')} \\
&= \frac{\pi_c |2\pi\Sigma_c|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_c)^T \Sigma_c^{-1}(\mathbf{x}-\mu_c)\right]}{\sum_{c'} \pi_{c'} |2\pi\Sigma_{c'}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T \Sigma_{c'}^{-1}(\mathbf{x}-\mu_{c'})\right]}
\end{aligned}
$$

- We will consider the form of this equation for various special cases:
- $\Sigma_1=\Sigma_0$,
- Σ_c tied, many classes
- General case

$\Sigma_1 = \Sigma_0$

- Class posterior simplifies to

$$
\begin{aligned}
p(Y=1|\mathbf{x}) &= \frac{p(\mathbf{x}|Y=1)p(Y=1)}{p(\mathbf{x}|Y=1)p(Y=1)+p(\mathbf{x}|Y=0)p(Y=0)} \\
&= \frac{\pi_1 \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\right]}{\pi_1 \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\right] + \pi_0 \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0)\right]} \\
&= \frac{\pi_1 e^{a_1}}{\pi_1 e^{a_1} + \pi_0 e^{a_0}} = \frac{1}{1+\frac{\pi_0}{\pi_1}e^{a_0-a_1}} \\
a_c &\stackrel{\text{def}}{=} -\frac{1}{2}(\mathbf{x}-\mu_c)^T\Sigma(\mathbf{x}-\mu_c)
\end{aligned}
$$

$\Sigma_1 = \Sigma_0$

- Class posterior simplifies to

$$
\begin{aligned}
p(Y=1|\mathbf{x}) &= \frac{1}{1+\exp\left[-\log\frac{\pi_1}{\pi_0} + a_0 - a_1\right]} \\
a_0 - a_1 &= -\tfrac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0) + \tfrac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1) \\
&= -(\mu_1-\mu_0)^T\Sigma^{-1}\mathbf{x} + \tfrac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)
\end{aligned}
$$

so

Linear function of x

$$
\begin{aligned}
p(Y=1|\mathbf{x}) &= \frac{1}{1+\exp\left[-\beta^T\mathbf{x}-\gamma\right]} = \sigma(\beta^T\mathbf{x}+\gamma) \\
\beta &\stackrel{\text{def}}{=} \Sigma^{-1}(\mu_1-\mu_0) \\
\gamma &\stackrel{\text{def}}{=} -\tfrac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0) + \log\frac{\pi_1}{\pi_0} \\
\sigma(\eta) &\stackrel{\text{def}}{=} \frac{1}{1+e^{-\eta}} = \frac{e^\eta}{e^\eta+1}
\end{aligned}
$$

Decision boundary

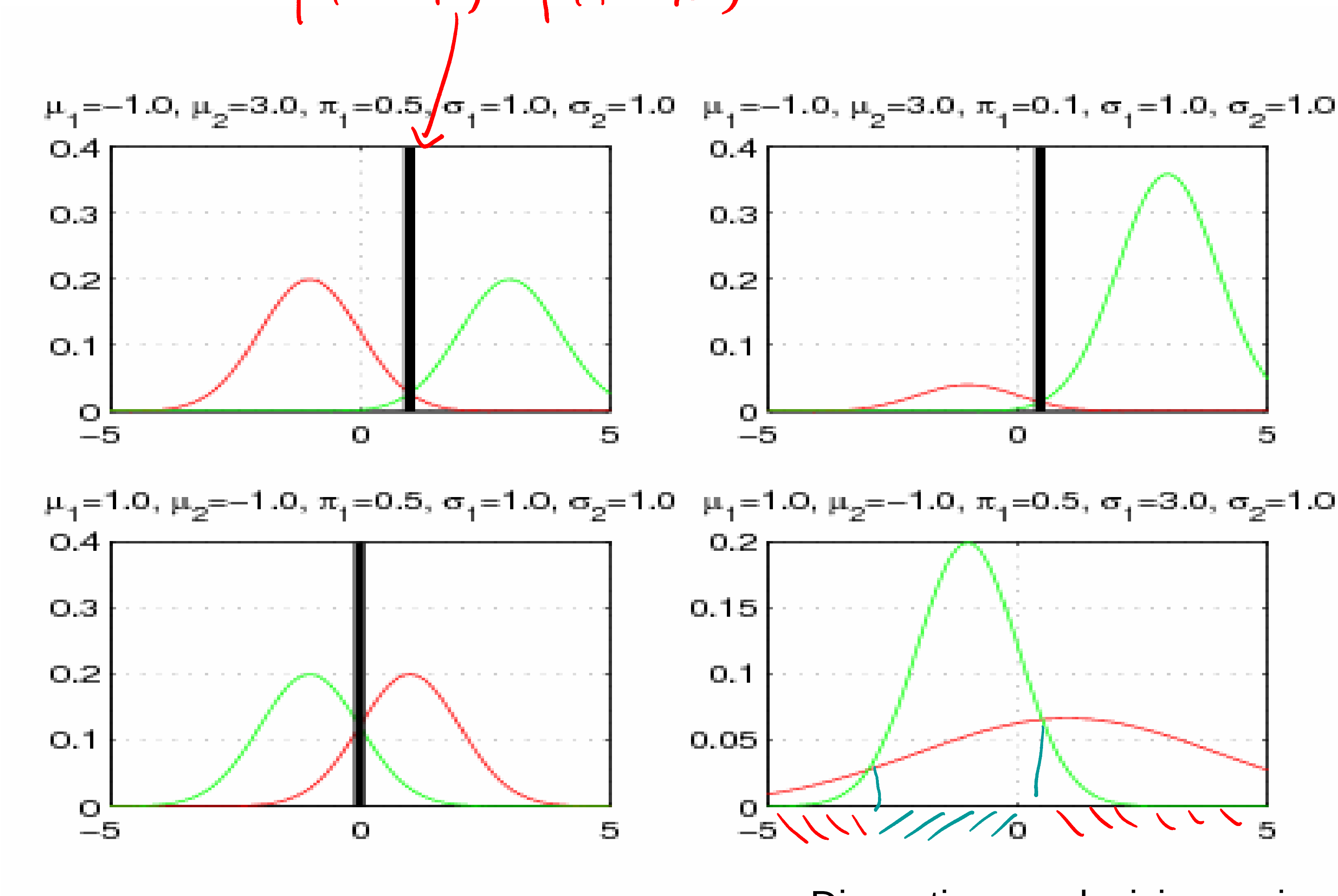
- Rewrite class posterior as

$$
\begin{aligned}
p(Y=1|\mathbf{x}) &= \sigma(\boldsymbol{\beta}^T\mathbf{x}+\gamma) = \sigma(\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0)) \\
\mathbf{w} &= \beta = \Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) \\
\mathbf{x}_0 &= -\frac{\gamma}{\boldsymbol{\beta}} = \tfrac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_0) - \frac{\log(\pi_1/\pi_0)}{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)
\end{aligned}
$$

- If $\Sigma = I$, then $w=(\mu_1-\mu_0)$ is in the direction of $\mu_1-\mu_0$, so the hyperplane is orthogonal to the line between the two means, and intersects it at x_0
- If $\pi_1=\pi_0$, then $x_0 = 0.5(\mu_1+\mu_0)$ is midway between the two means
- If π_1 increases, x_0 decreases, so the boundary shifts toward μ_0 (so more space gets mapped to class 1)

Decision boundary in 1d

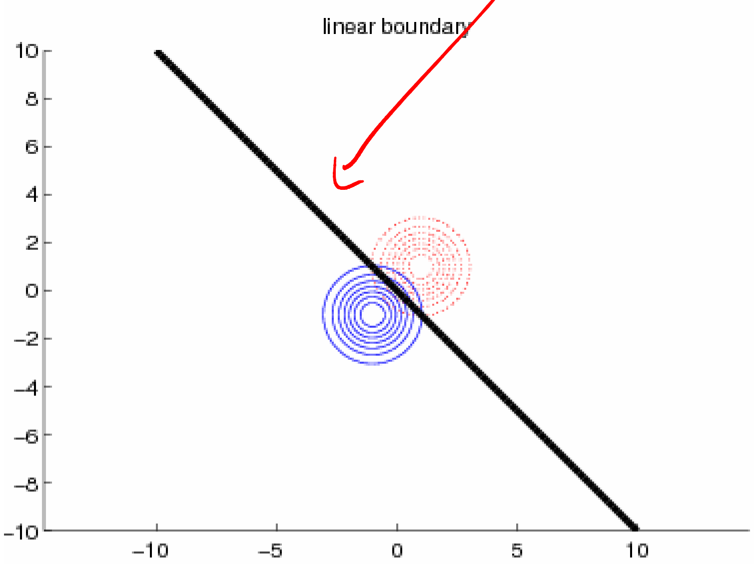
$p(y=1|x) = p(y=0|x)$

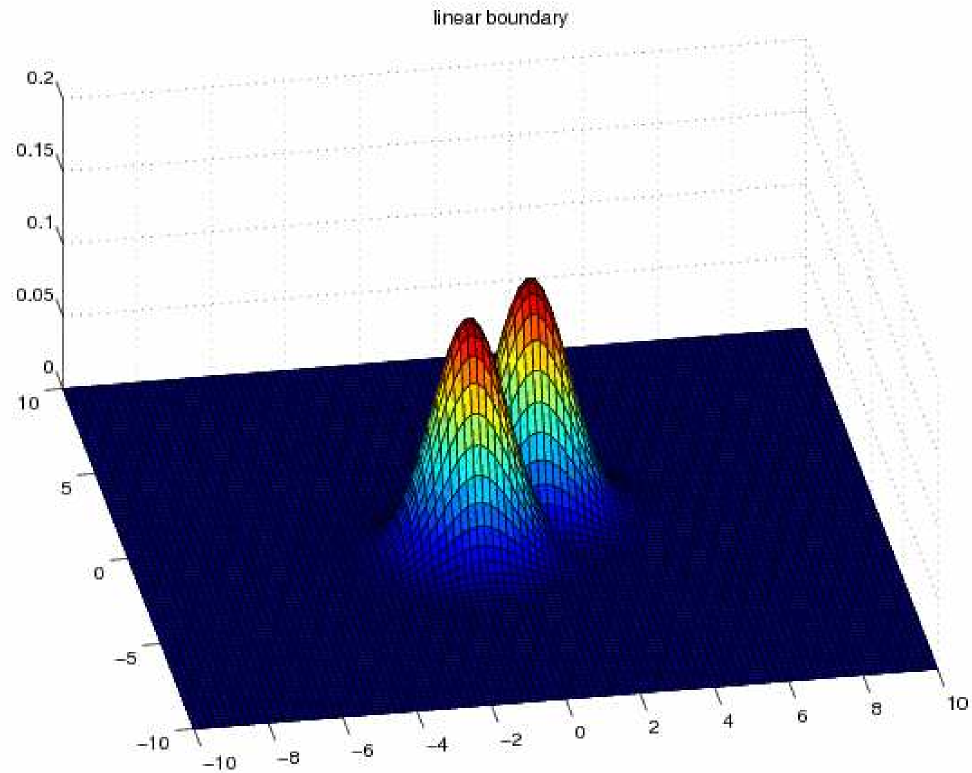


Discontinuous decision region

Decision boundary in 2d

$p(y=1|x) = p(y=0|x)$





Tied Σ, many classes

- Similarly to before

$$
\begin{aligned}
p(Y=c|\mathbf{x}) &= \frac{\pi_c \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_c)^T \Sigma_c^{-1} (\mathbf{x}-\mu_c)\right]}{\sum_{c'} \pi_{c'} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_{c'})^T \Sigma_{c'}^{-1} (\mathbf{x}-\mu_{c'})\right]} \\
&= \frac{\exp\left[\mu_{'}^T \Sigma^{-1} x - \frac{1}{2}\mu_c^T \Sigma^{-1} \mu_c + \log \pi_c\right]}{\sum_{c'} \exp\left[\mu_{c'}^T \Sigma^{-1} \mathbf{x} - \frac{1}{2}\mu_{c'}^T \Sigma^{-1} \mu_{c'} + \log \pi_{c'}\right]} \\
\theta_c &\stackrel{\text{def}}{=} \begin{pmatrix} -\mu_c^T \Sigma^{-1} \mu_c + \log \pi_c \\ \Sigma^{-1} \mu_c \end{pmatrix} = \begin{pmatrix} \gamma_c \\ \beta_c \end{pmatrix} \\
p(Y=c|\mathbf{x}) &= \frac{e^{\theta_c^T \mathbf{x}}}{\sum_{c'} e^{\theta_{c'}^T \mathbf{x}}} = \frac{e^{\beta_c^T \mathbf{x} + \gamma_c}}{\sum_{c'} e^{\beta_{c'}^T \mathbf{x} + \gamma_{c'}}}
\end{aligned}
$$

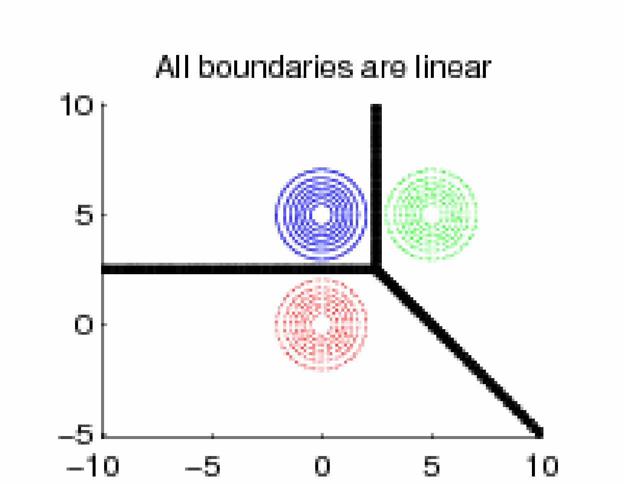
- This is the multinomial logit or softmax function

Tied Σ, many classes

- Discriminant function

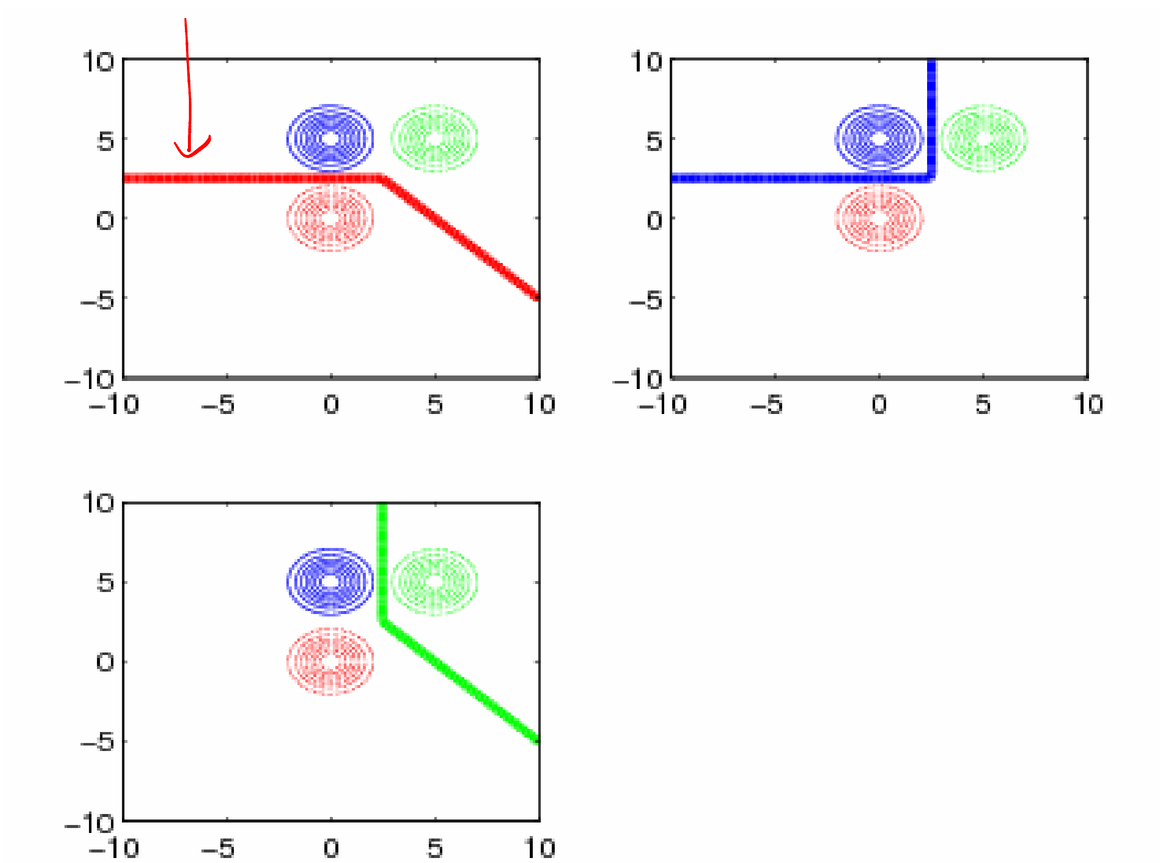
$$
\begin{aligned}
g_c(\mathbf{x}) &= -\tfrac{1}{2}(\mathbf{x}-\mu_c)^T\Sigma^{-1}(\mathbf{x}-\mu_c) + \log p(Y=c) = \beta_c^T\mathbf{x} + \beta_{c0} \\
\beta_c &= \Sigma^{-1}\mu_c \\
\beta_{c0} &= -\tfrac{1}{2}\mu_c^T\Sigma^{-1}\mu_c + \log \pi_c
\end{aligned}
$$

- Decision boundary is again linear, since $x^T \Sigma x$ terms cancel
- If $\Sigma = I$, then the decision boundaries are orthogonal to $\mu_i - \mu_j$, otherwise skewed



Decision boundaries

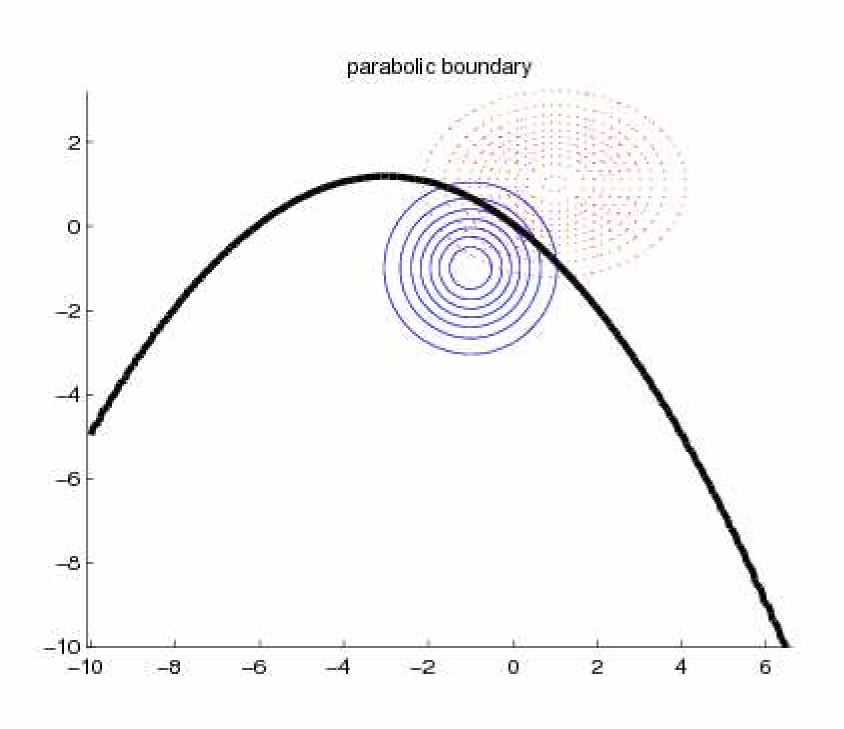
$$g_1(x) - \max\left(g_2(x),\ g_3(x)\right) = 0$$

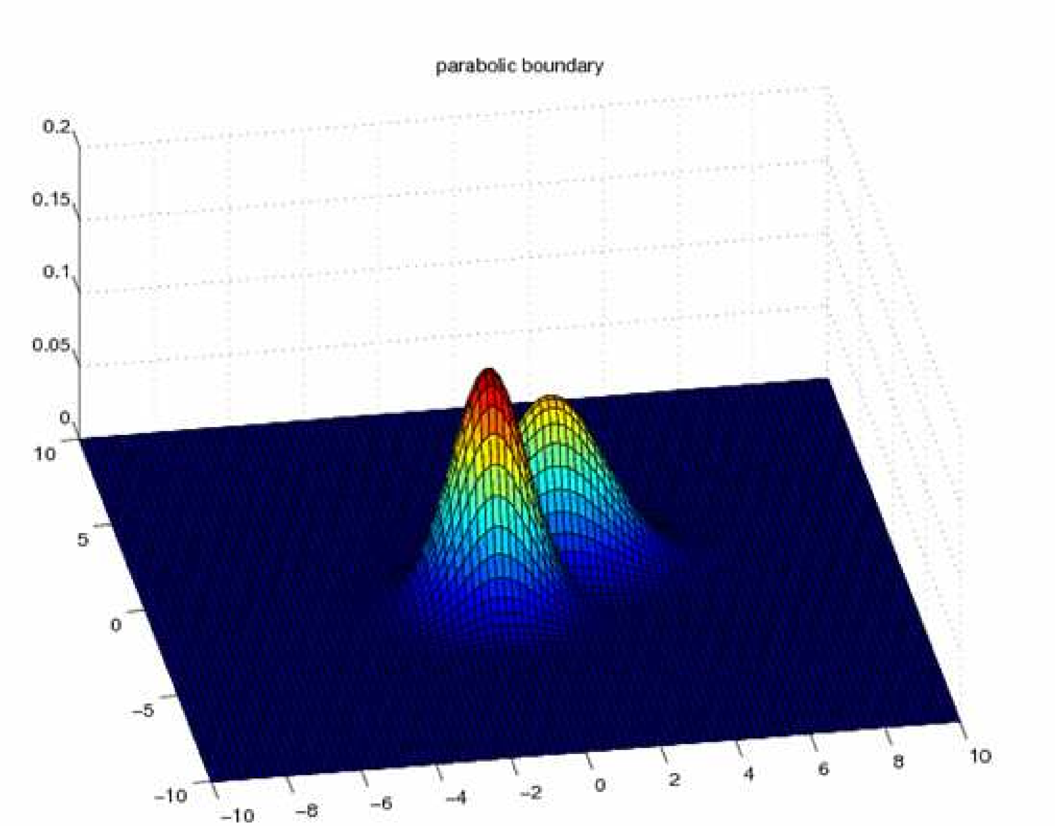



```
[x,y] = meshgrid(linspace(-10,10,100), linspace(-10,10,100));
g1 = reshape(mvnpdf(X, mu1(:)', S1), [m n]); ...
contour(x,y,g2*p2-max(g1*p1, g3*p3),[0 0],'-k');
```

Σ_0, Σ_1 arbitrary

- If the Σ are unconstrained, we end up with cross product terms, leading to quadratic decision boundaries



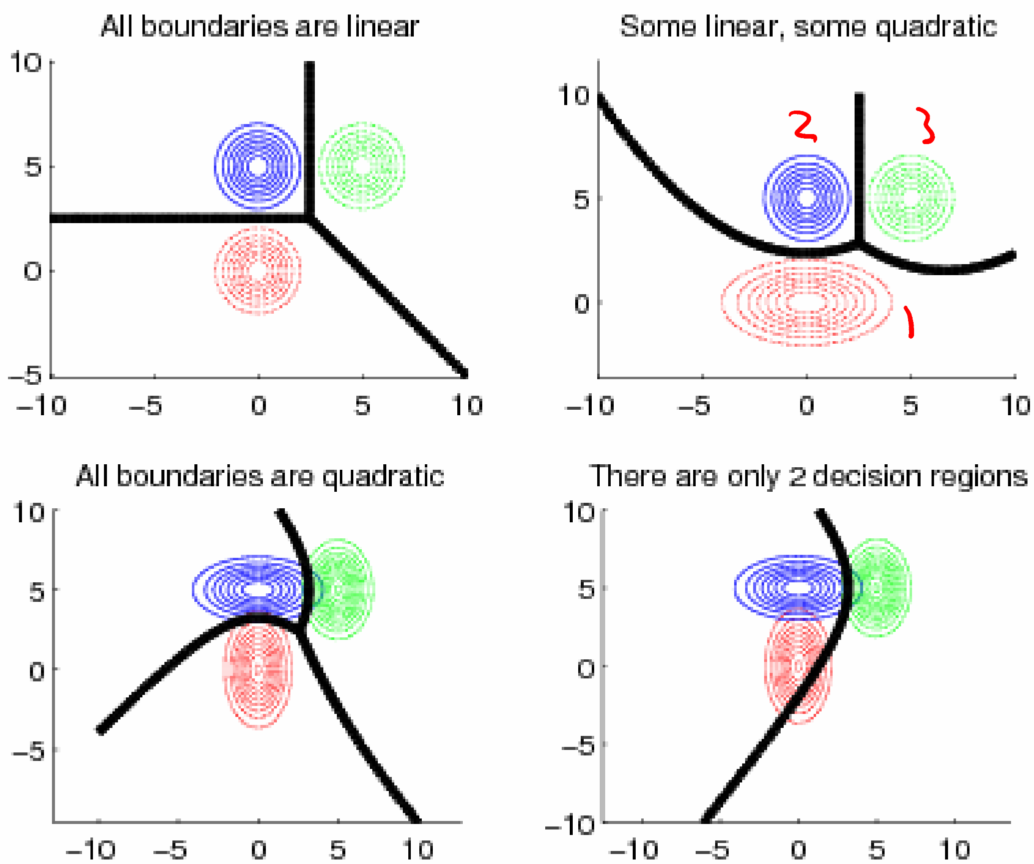


General case

$$\mu_1 = (0,0), \mu_2 = (0,5), \mu_3 = (5,5), \pi = (1/3, 1/3, 1/3)$$

$\Sigma_c = I$

$\Sigma_1 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_2 = \Sigma_3 = I$



$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$

$\Sigma_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$

$\Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$

$\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$

$\Sigma_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$

$\Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$

$\pi = (0, 1/2, 1/2)$