CS 340 Lec. 18: Multivariate Gaussian Distributions and Linear Discriminant Analysis

AD

March 2011

Multivariate Gaussian

- Consider data $\left\{\mathbf{x}^i\right\}_{i=1}^N$ where $\mathbf{x}^i \in \mathbb{R}^D$ and we assume they are independent and identically distributed.
- A standard pdf used to model multivariate real data is the multivariate Gaussian or normal

$$\begin{aligned} p\left(\mathbf{x}|\,\boldsymbol{\mu}, \Sigma\right) &= \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \Sigma\right) \\ &= \frac{1}{(2\pi)^{D/2}\left|\Sigma\right|^{1/2}} \exp\Big(-\frac{1}{2}\underbrace{\left(\mathbf{x}-\boldsymbol{\mu}\right)^{\mathsf{T}}\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)}_{\text{Mahalanobis distance}}\Big). \end{aligned}$$

- It can be shown that $\boldsymbol{\mu}$ is the mean and Σ is the covariance of $\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \Sigma\right)$; i.e.

$$\mathbb{E}\left(\mathbf{X}\right) = \boldsymbol{\mu} \text{ and } cov\left(\mathbf{X}\right) = \Sigma.$$

- It will be used extensively in our discussion on unsupervised learning and can also used for generative classifiers (i.e. discriminant analysis).

Special Cases

- When $D = 1$, we are back to

$$p\left(x|\,\mu,\sigma^2\right) = \mathcal{N}\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\left(x-\mu\right)^2\right)$$

- When $D = 2$ and writing

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

where $\rho = corr\left(X_1, X_2\right) \in [-1,1]$ we have

$$p\left(\mathbf{x}|\,\boldsymbol{\mu},\Sigma\right) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}$$
$$\times \exp\left(-\frac{1}{2(1-\rho^2)}\left\{\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right\}\right)$$

Graphical Illustrations

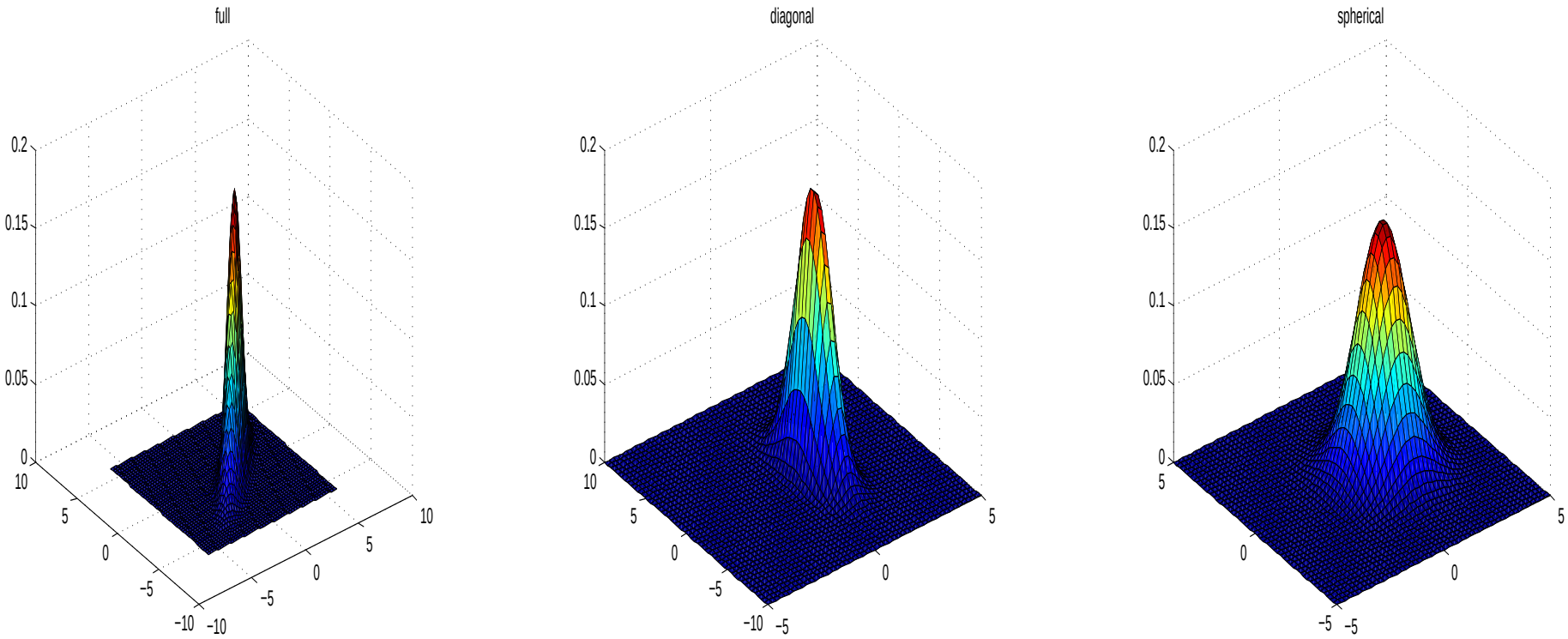


Illustration of 2D Gaussian pdfs for different covariance matrices (left): full, (middle): diagonal, (right): spherical.

Why are the contours of a multivariate Gaussian elliptical

- If we plot the values of $\mathbf{x}$ s.t. $p(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ is equal to a constant, i.e. s.t. $(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = c > 0$ where c is given, then we obtain an ellipse.
- Σ is a positive definite matrix so we have

$$\Sigma = U \Lambda U^{\mathsf{T}}$$

where U is an orthonormal matrix of eigenvectors, i.e. $U^{\mathsf{T}} U = I$, and $\Lambda = diag(\lambda_1, ..., \lambda_D)$ with $\lambda_k \geq 0$ is the diagonal matrix of eigenvalues.

- Hence, we have

$$\Sigma^{-1} = \left(U^{\mathsf{T}}\right)^{-1} \Lambda^{-1} U^{-1} = U \Lambda^{-1} U^{\mathsf{T}} = \sum_{k=1}^{D} \frac{1}{\lambda_k} \mathbf{u}_k \mathbf{u}_k^{\mathsf{T}},$$

so

$$(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = \sum_{k=1}^{D} \frac{y_k^2}{\lambda_k} \text{ where } y_k = \mathbf{u}_k^{\mathsf{T}} (\mathbf{x}-\boldsymbol{\mu}).$$

Graphical Illustrations

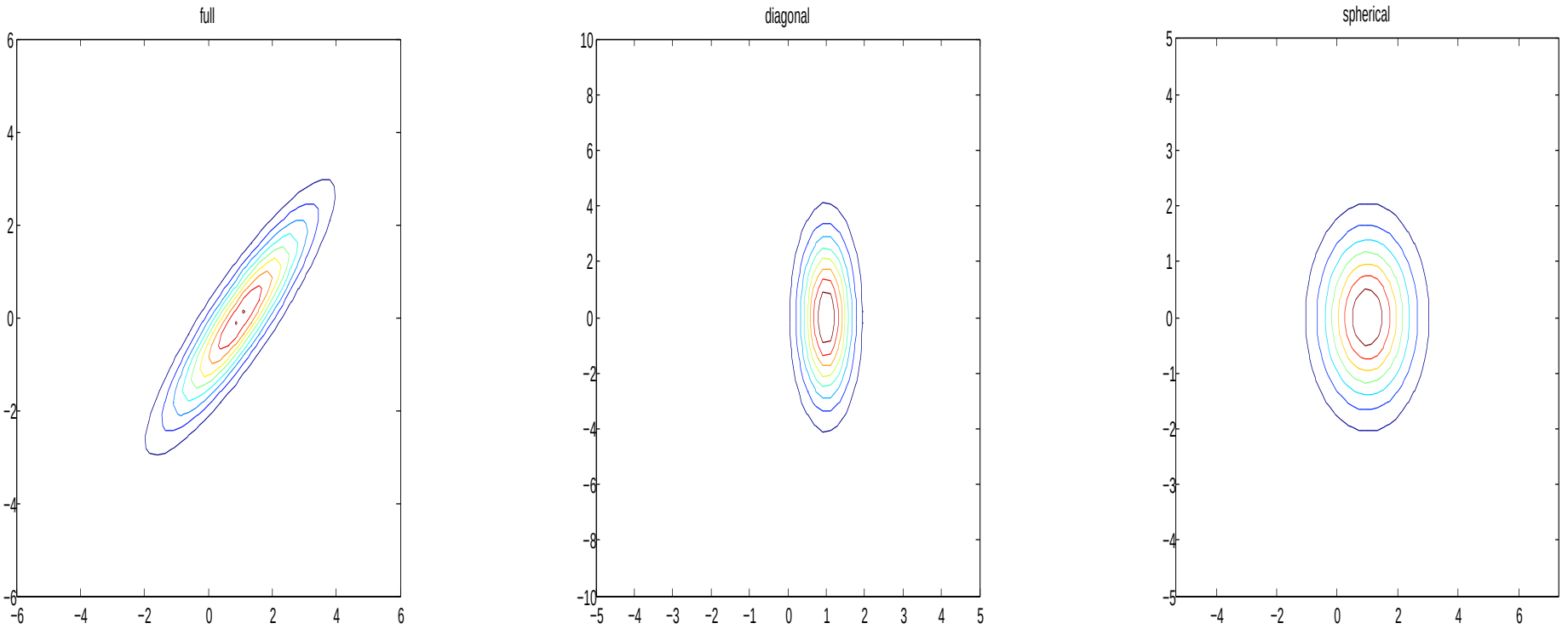


Illustration of 2D Gaussian pdfs level sets for different covariance matrices (left): full, (middle): diagonal, (right): spherical.

Properties of Multivariate Gaussians

- Marginalization is straightforward.
- Conditioning is easy; e.g. if $\mathbf{X} = (\mathbf{X}_1 \ \mathbf{X}_2)$ with

$$p(\mathbf{x}) = p(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

then

$$p(\mathbf{x}_1 | \mathbf{x}_2) = \frac{p(\mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{x}_2)} = \mathcal{N}\left(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \Sigma_{1|2}\right)$$

with

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Independence and Correlation for Gaussian Variables

- It is well-known that independence implies uncorrelations; i.e. if the components $(X_1, ..., X_D)$ of a vector $\mathbf{X}$ are independent then they are uncorrelated. However, uncorrelated does not imply independence in the general case.
- If the components $(X_1, ..., X_D)$ of a vector $\mathbf{X}$ distributed according to a multivariate Gaussian are uncorrelated then they are independent.
- **Proof**. If $(X_1, ..., X_D)$ are uncorrelated then $\Sigma = diag\left(\sigma_1^2, ..., \sigma_D^2\right)$ and $|\Sigma| = \prod_{k=1}^{D} \sigma_k^2$ so

$$p\left(\mathbf{x}|\,\boldsymbol{\mu}, \Sigma\right) = \prod_{k=1}^{D} p\left(x_k|\,\mu_k, \sigma_k^2\right) = \prod_{k=1}^{D} \mathcal{N}\left(x_k; \mu_k, \sigma_k^2\right)$$

ML Parameter Learning for Multivariate Gaussian

- Consider data $\left\{\mathbf{x}^i\right\}_{i=1}^N$ where $\mathbf{x}^i \in \mathbb{R}^D$ and assume they are independent and identically distributed from $\mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}, \Sigma\right)$.
- The ML parameter estimates of $(\boldsymbol{\mu}, \Sigma)$ maximize by definition

$$\begin{aligned} & \sum_{i=1}^N \log\ \mathcal{N}\left(\mathbf{x}^i; \boldsymbol{\mu}, \Sigma\right) \\ = \quad & -\frac{ND}{2}\log\left(2\pi\right) - \frac{N}{2}\log\left|\Sigma\right| - \frac{1}{2}\sum_{i=1}^N \left(\mathbf{x}^i - \boldsymbol{\mu}\right)^{\mathsf{T}} \Sigma^{-1}\left(\mathbf{x}^i - \boldsymbol{\mu}\right). \end{aligned}$$

- We obtain after painful calculations the fairly intuitive results

$$\widehat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^N \mathbf{x}^i}{N},\ \widehat{\Sigma} = \frac{\sum_{i=1}^N \left(\mathbf{x}^i - \widehat{\boldsymbol{\mu}}\right)\left(\mathbf{x}^i - \widehat{\boldsymbol{\mu}}\right)^{\mathsf{T}}}{N}.$$

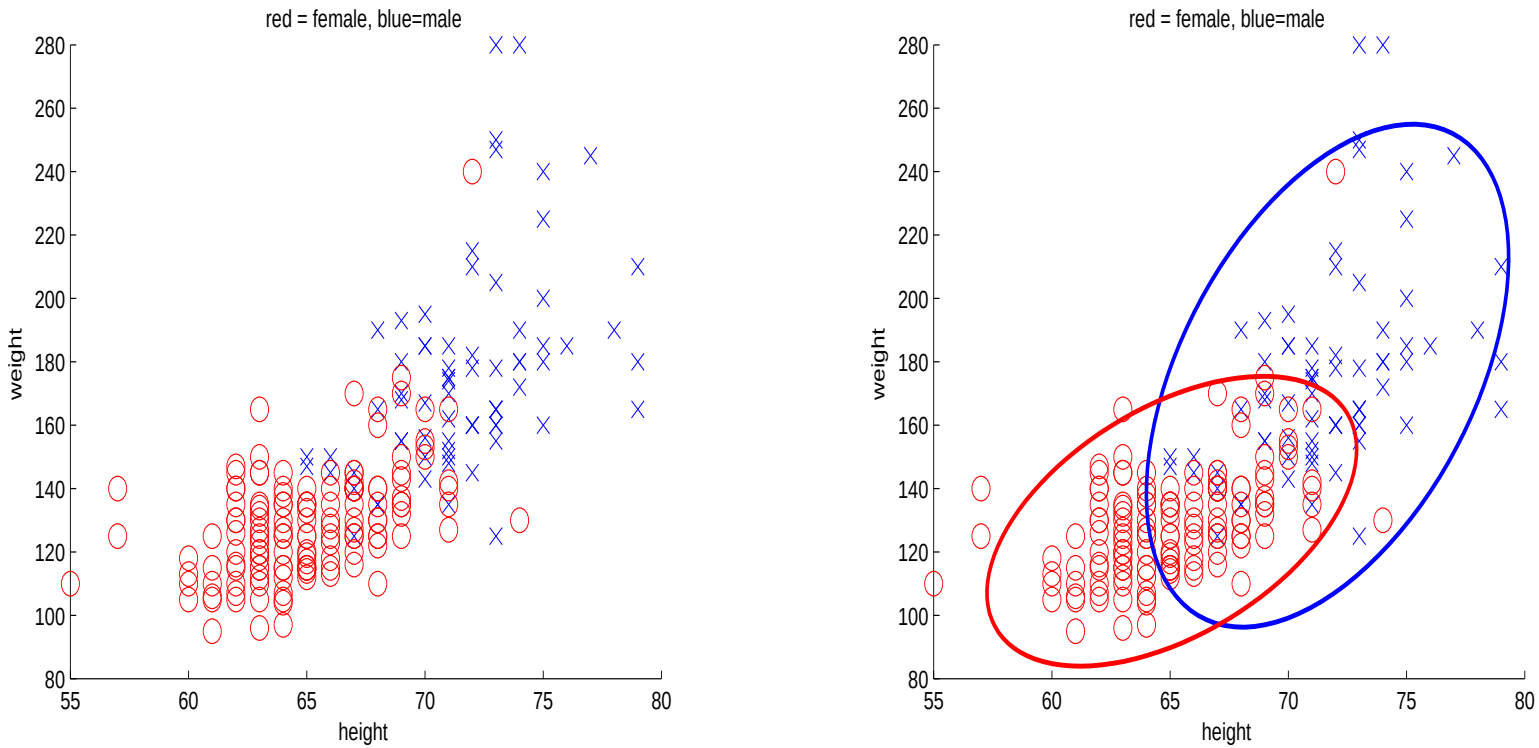
Application to Supervised Learning using Bayes Classifier

- Assume you are given some training data $\left\{\mathbf{x}^i, y^i\right\}_{i=1}^N$ where $\mathbf{x}^i \in \mathbb{R}^D$ and $y^i \in \{1, 2, ..., C\}$ can take C different values.
- Given an input test data $\mathbf{x}$, you want to predict/estimate the output y associated to $\mathbf{x}$.
- Previously we have followed a probabilistic approach

$$p\left(y = c \middle| \mathbf{x}\right) = \frac{p\left(y = c\right) p\left(\mathbf{x} \middle| y = c\right)}{\sum_{c'=1}^{C} p\left(y = c'\right) p\left(\mathbf{x} \middle| y = c'\right)}.$$

- This requires modelling and learning the parameters of the class conditional density of features $p\left(\mathbf{x} \middle| y = c\right)$.

Height Weight Data



(left) Height/Weight data for female/male (right) 2d Gaussians fit to each class. 95% of the proba mass is inside the ellipse

Supervised Learning using Bayes Classifier

- Assume we pick

$$p(\mathbf{x}|\, y = c) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_c, \Sigma_c)$$

and $p(y = c) = \pi_c$ then

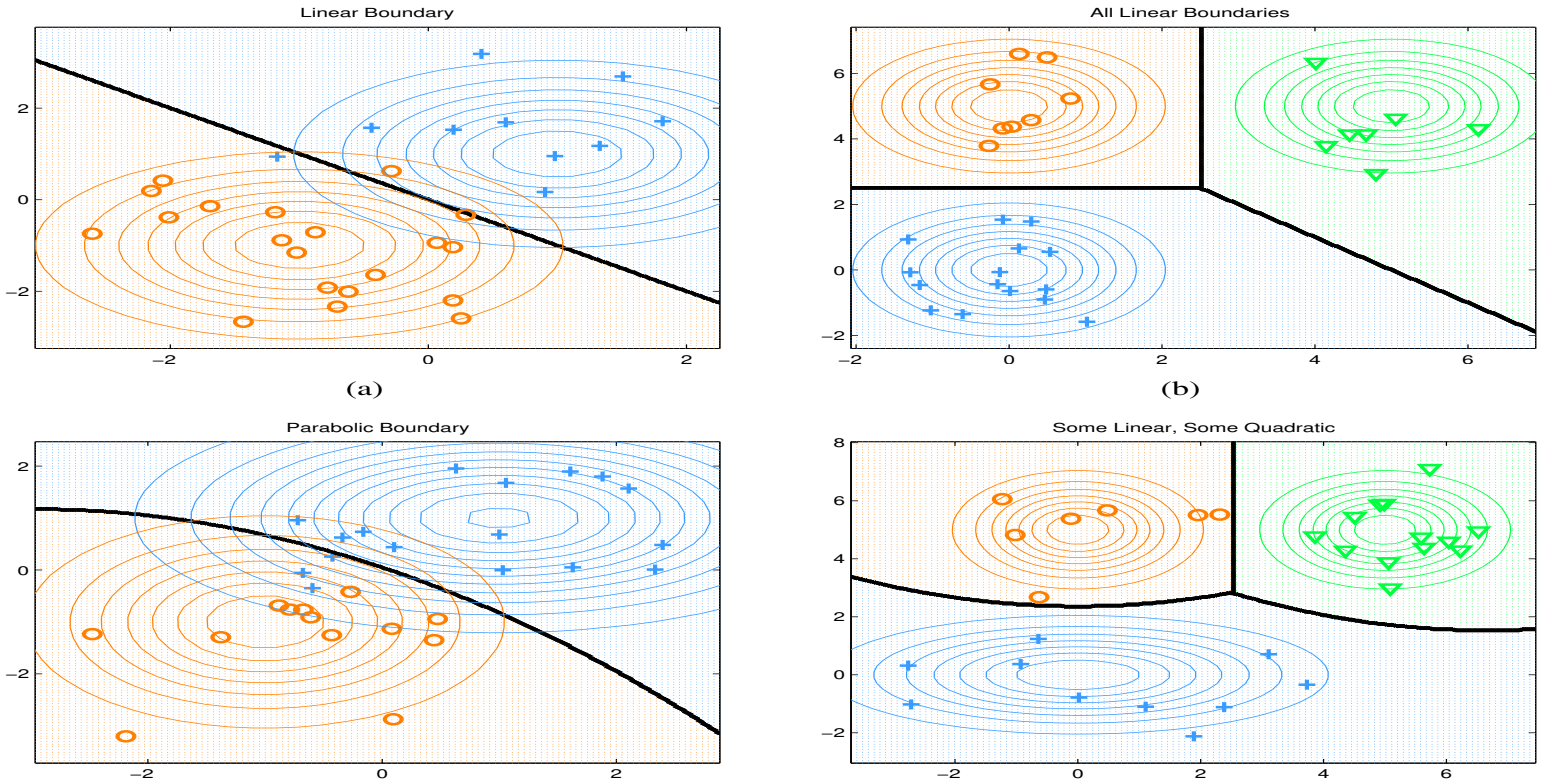
$$\begin{aligned} p(y = c|\, \mathbf{x}) &\propto \pi_c \left|\Sigma_c\right|^{-1/2} \exp(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^{\mathsf{T}} \Sigma_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)) \\ &= \exp(\boldsymbol{\mu}_c^{\mathsf{T}} \Sigma_c^{-1} \mathbf{x} - \tfrac{1}{2}\boldsymbol{\mu}_c^{\mathsf{T}} \Sigma_c^{-1} \boldsymbol{\mu}_c + \log \pi_c) \exp\left(-\tfrac{1}{2}\mathbf{x}^{\mathsf{T}} \Sigma_c^{-1} \mathbf{x}\right) \end{aligned}$$

- For models where $\Sigma_c = \Sigma$ then this is known as *linear discriminant analysis*

$$p(y = c|\, \mathbf{x}) = \frac{\exp\left(\beta_c^{\mathsf{T}} \mathbf{x} + \gamma_c\right)}{\sum_{c'=1}^{C} \exp\left(\beta_{c'}^{\mathsf{T}} \mathbf{x} + \gamma_{c'}\right)}$$

where $\beta_c = \Sigma^{-1} \boldsymbol{\mu}_c$, $\gamma_c = -\frac{1}{2}\boldsymbol{\mu}_c^{\mathsf{T}} \Sigma^{-1} \boldsymbol{\mu}_c + \log \pi_c$ and the model is very similar to logistic regression.

Decision Boundaries



Decision boundaries in 2D for 2 and 3 class case.

Binary Classification

- Consider the case where $C = 2$ then one can check that

$$p(y = 1|\mathbf{x}) = g\left((\beta_1 - \beta_0)^{\mathsf{T}}\mathbf{x} + \gamma_1 - \gamma_0\right)$$

- We have

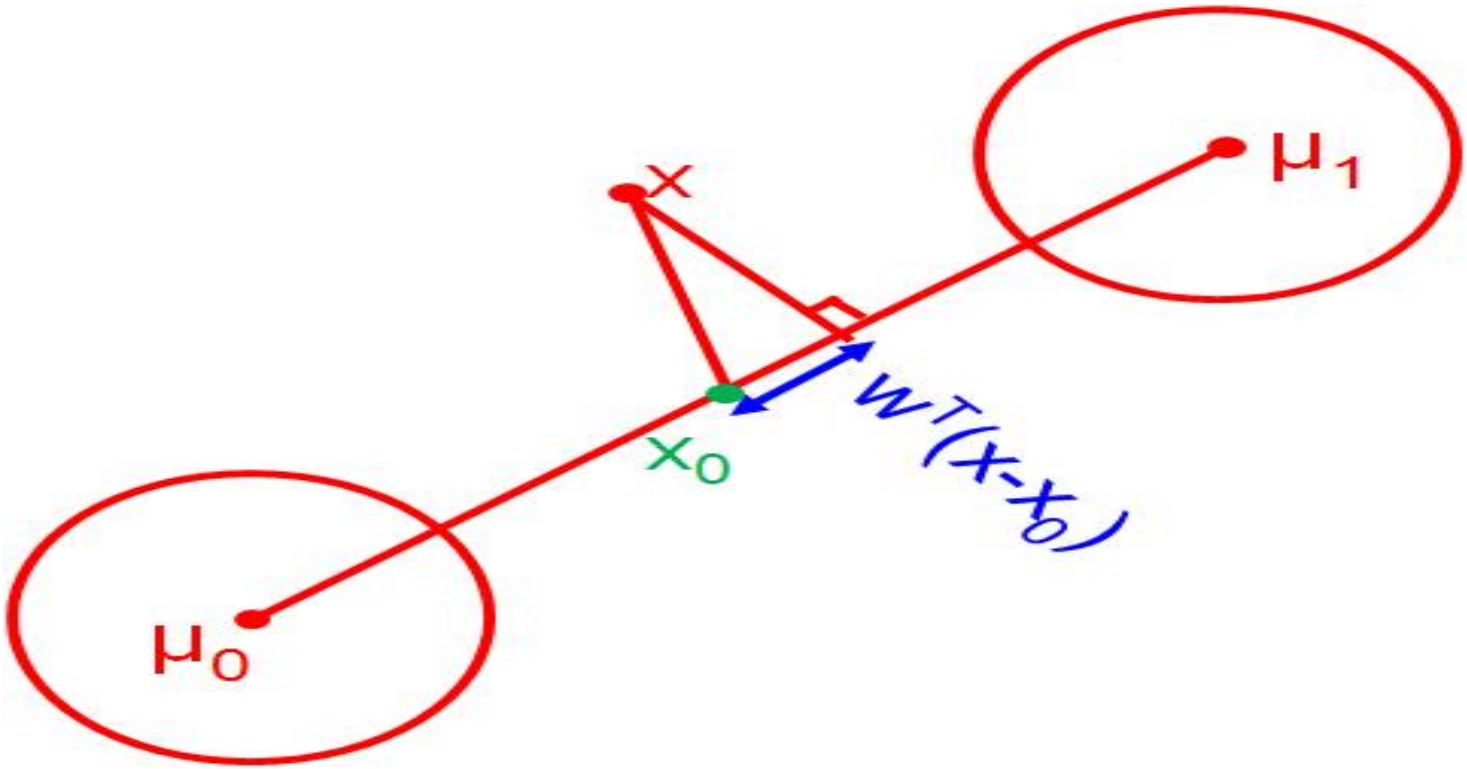
$$\begin{aligned}
\gamma_1 - \gamma_0 &= -\frac{1}{2}(\mu_1 - \mu_0)^{\mathsf{T}}\Sigma^{-1}(\mu_1 + \mu_0) + \log(\pi_1/\pi_0), \\
\mathbf{x}_0 &:= \frac{1}{2}(\mu_1 + \mu_0) - \frac{(\mu_1 - \mu_0)\log(\pi_1/\pi_0)}{(\mu_1 - \mu_0)^{\mathsf{T}}\Sigma^{-1}(\mu_1 - \mu_0)} \\
\mathbf{w} &:= \beta_1 - \beta_0 = \Sigma^{-1}(\mu_1 - \mu_0)
\end{aligned}$$

  then

$$p(y = 1|\mathbf{x}) = g\left(\mathbf{w}^{\mathsf{T}}(\mathbf{x} - \mathbf{x}_0)\right)$$

- $\mathbf{x}$ is shifted by $\mathbf{x}_0$ and then projected onto the line $\mathbf{w}$.

Binary Classification

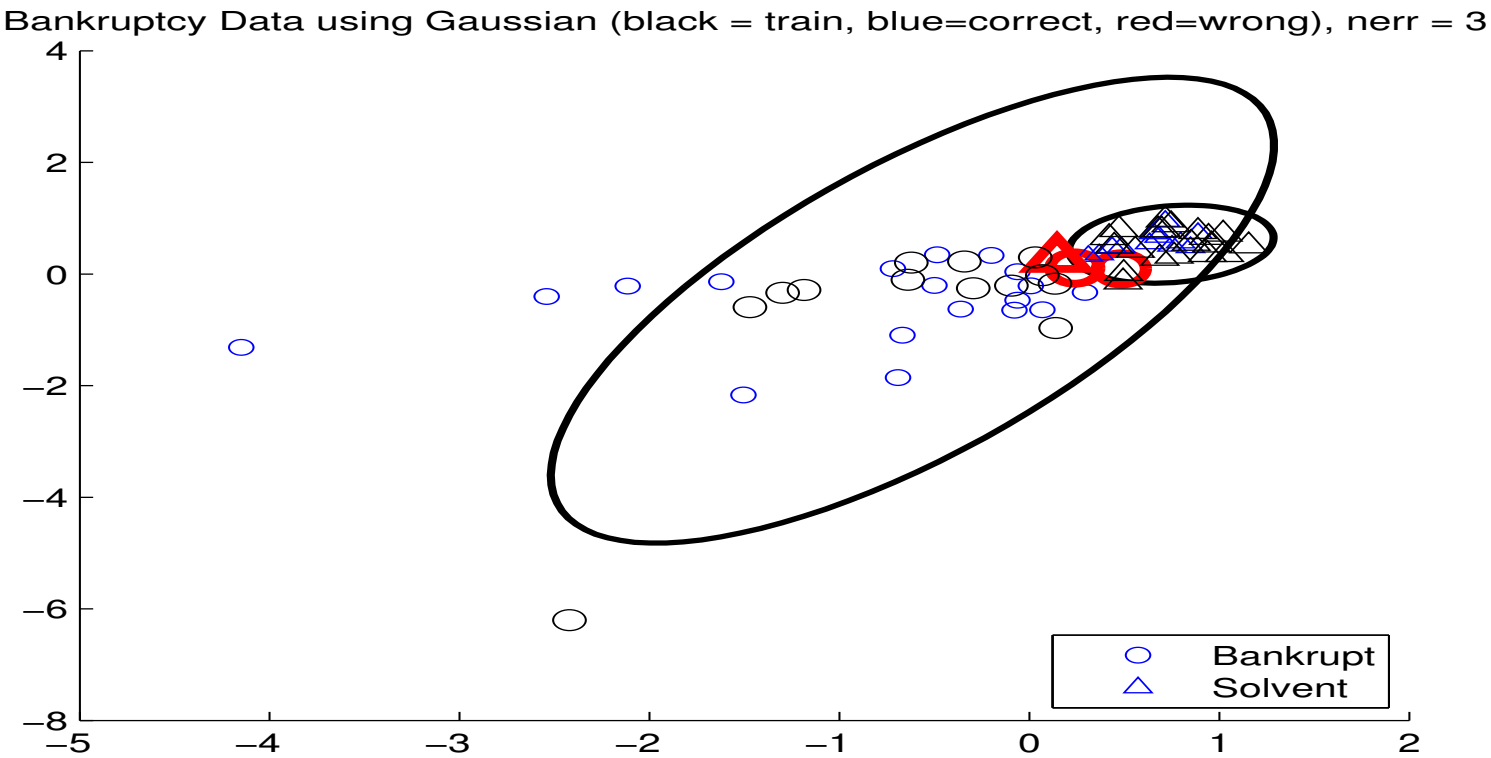


Example where $\Sigma = \sigma^2 I$ so **w** is in the direction of $\mu_1 - \mu_0$.

Generative or Discriminative Classifiers

- When the model for class conditional densities is correct, resp. incorrect, generative classifiers will typically outperform, resp. underperform, discriminative classifiers for large enough datasets.
- Generative classifiers can be difficult to learn whereas Discriminative classifiers try to learn directly the posterior probability of interest.
- Generative classifiers can handle missing data easily, discriminative methods cannot.
- Discriminative can be more flexible; e.g. substitute $\mathbf{x}$ to $\Phi(\mathbf{x})$.

Application



Discriminant analysis on the bankruptcy data set. Gaussian class conditional densities. Estimated labels, based on the posterior proba of belonging to each class, are computed. If incorrect, the point is colored read, otherwise in blue (Training data are black).