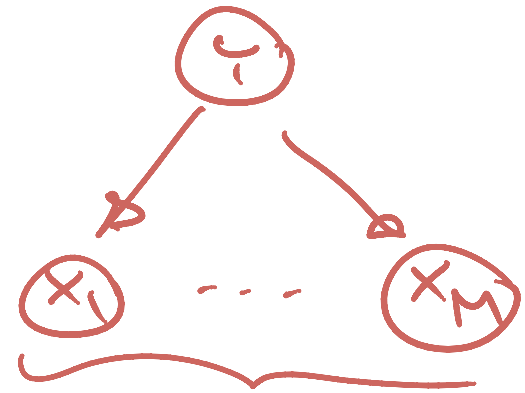
Review

- Train-test split
- Cross-validation
- Regularization and model complexity
  - L_1, L_2

Review



DIFFERENT FEATURES OF THE SAME EXAMPLE

- Classification w/ Naïve Bayes
- Features assumed independent given class
- Prediction: y = $\arg\max_y \left[p(y) \prod_j (x_j | y) \right]$
- Variations: Bag of Words, Gaussian NB

$$P(x_j \mid y=1) = \theta_{1j}^{x_j}(1-\theta_{1j})^{(1-x_j)}$$
$$\quad\quad\; y=0 \;\; = \theta_{0j} \ldots$$

A closer look at NB

- Y = 1 if: $P(Y=1)\prod_j P(X_j \mid Y=1) \geq P(Y=0)\prod_j P(X_j \mid Y=0)$

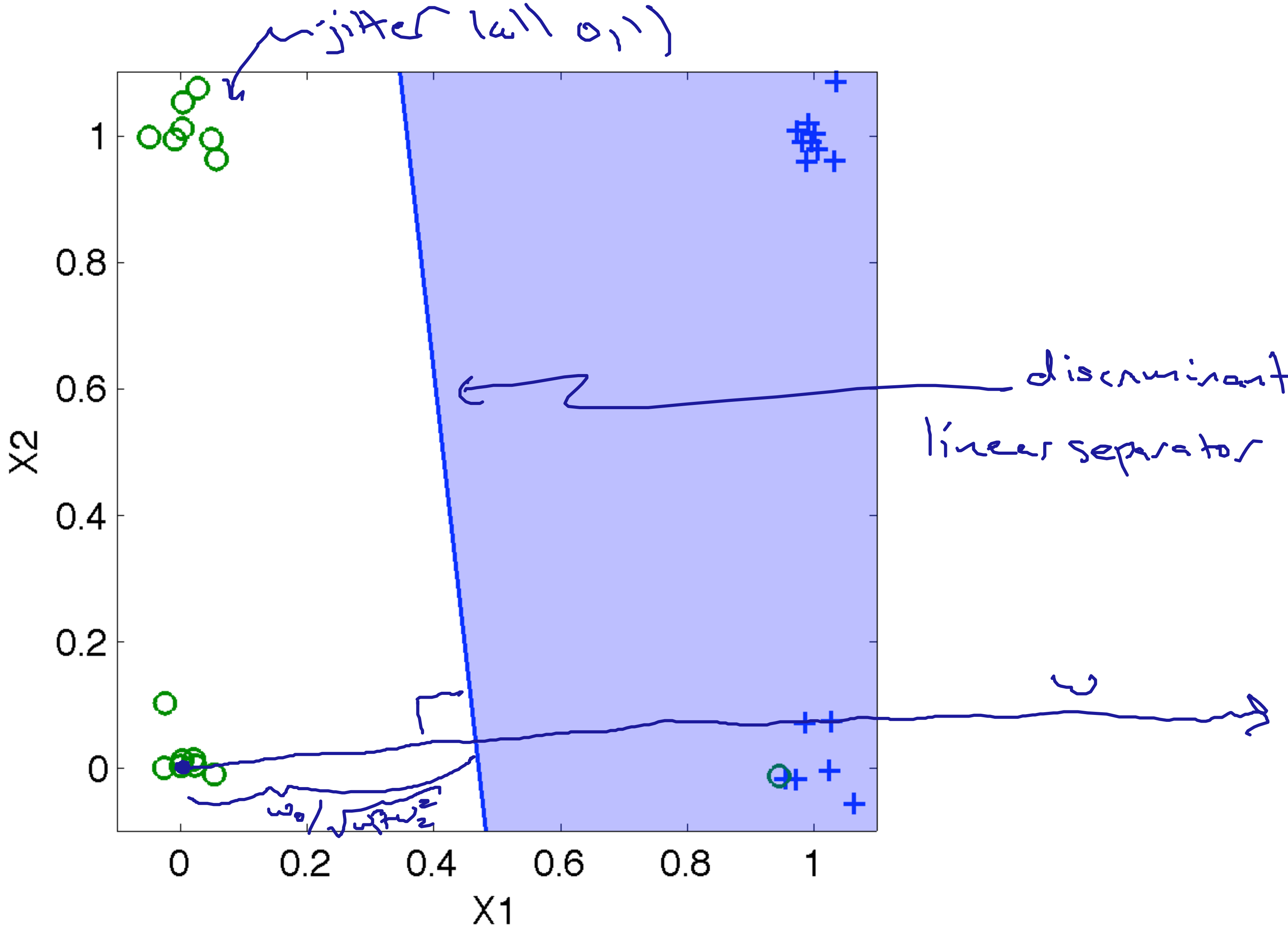
$$\log(P(Y=1)) - \log(P(Y=0)) + \sum_j \underbrace{\left[\log P(X_j \mid Y=1) - \log P(X_j \mid Y=0)\right]}_{x_j \log(\theta_{1j}) + (1-x_j)\log(1-\theta_{1j}) - x_j \log \theta_{0j} - (1-x_j)\log(1-\theta_{0j})} \geq 0$$

$$w_0 + \sum_j w_j x_j \geq 0$$

$$w_0 = \log P(Y=1) - \log P(Y=0) + \sum_j \log(1-\theta_{1j}) - \sum_j \log(1-\theta_{0j})$$

$$w_j = \log(\theta_{1j}) - \log(1-\theta_{1j}) - \log(\theta_{0j}) + \log(1-\theta_{0j})$$

Linear discriminants

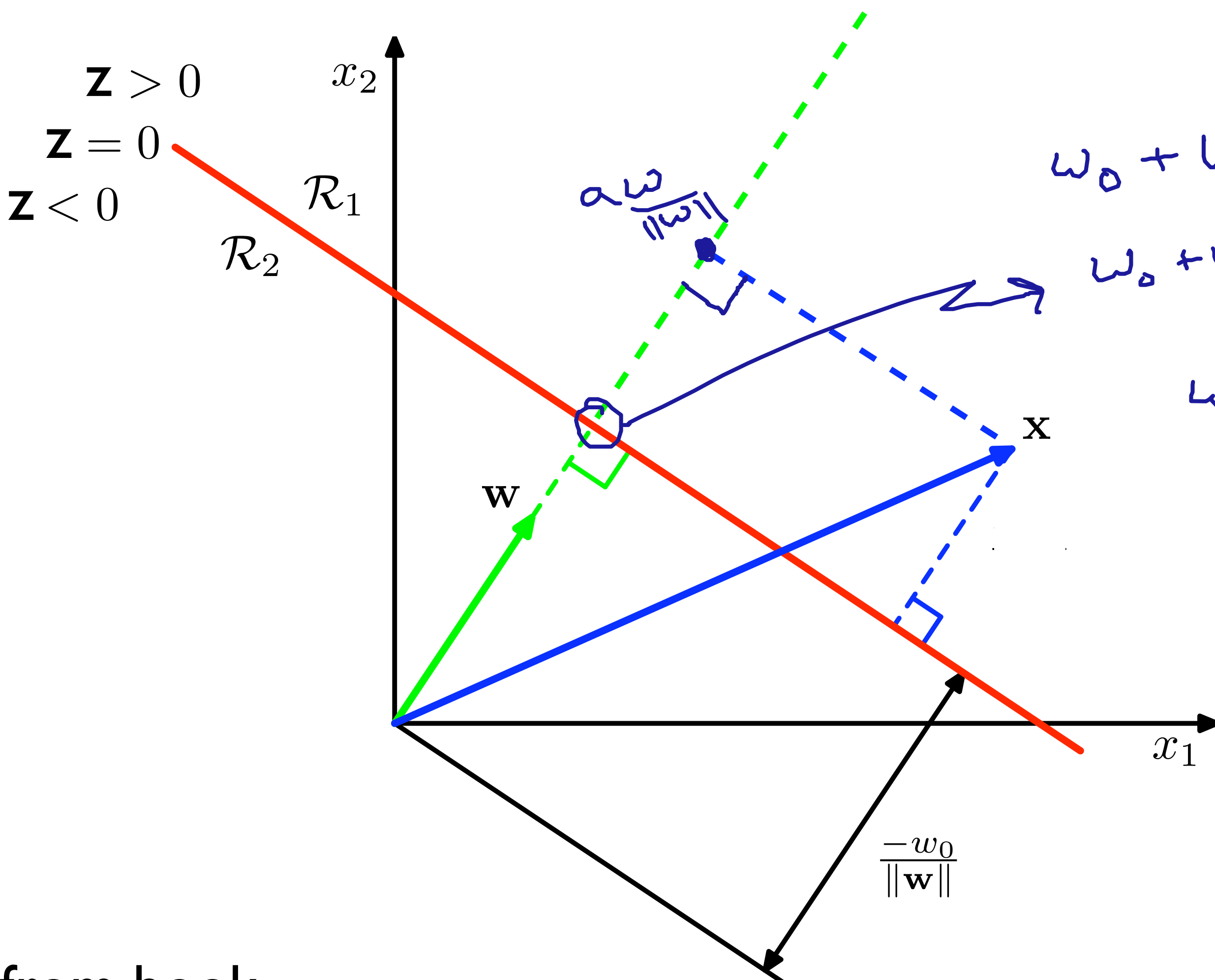


Geometry of a discriminant

$a \frac{w}{\|w\|} \cdot w = x \cdot w$

$a \frac{\|w\|^2}{\|w\|} = x \cdot w$

$\underline{a = x \cdot w / \|w\|}$



$w_0 + w^T x = 0$

$w_0 + w_1 x_1 + w_2 x_2 = 0$

$w_0 + w^T b \frac{w}{\|w\|} = 0$

$w_0 + b \|w\| = 0$

$\frac{-w_0}{\|w\|} = b$

figure adapted from book

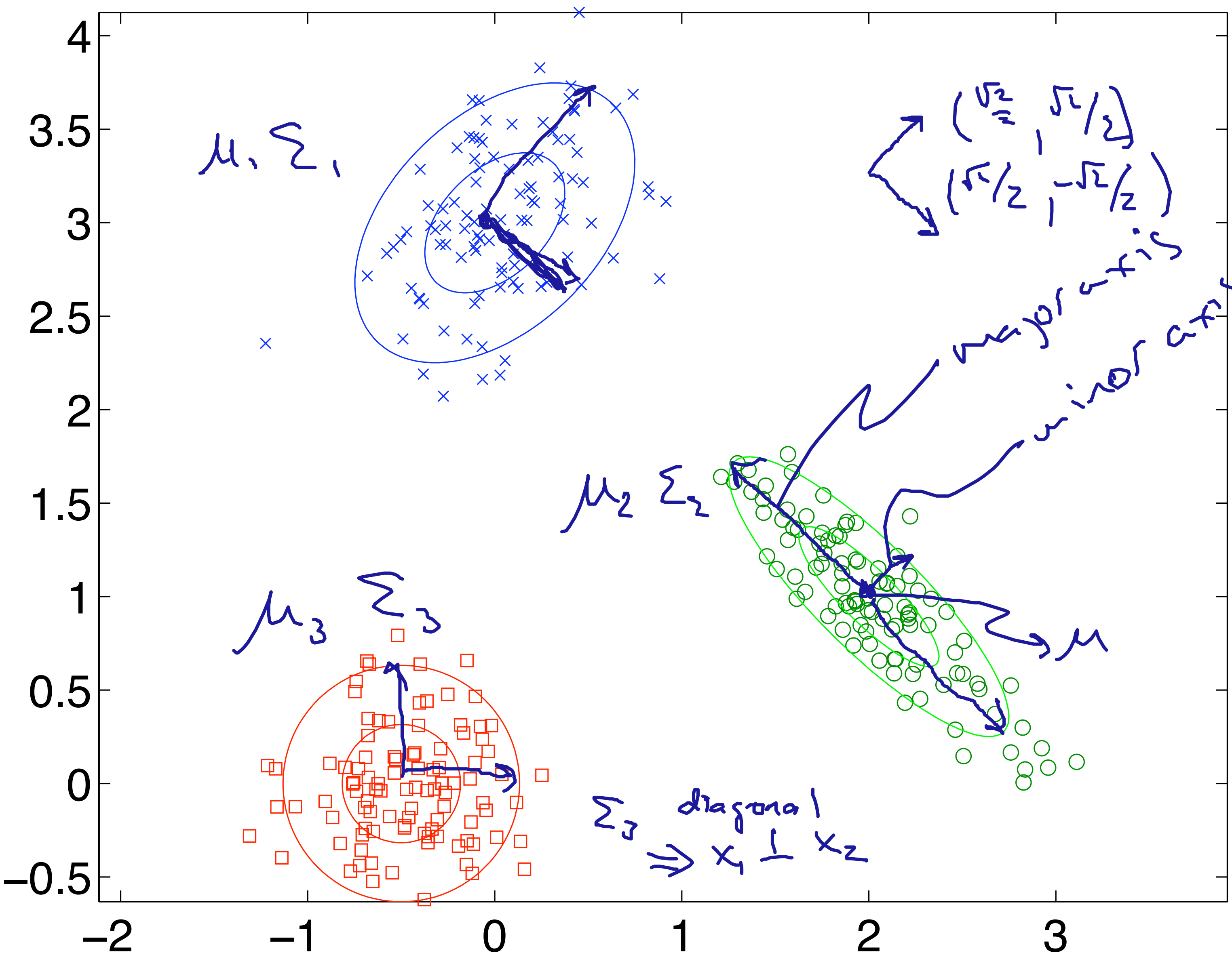
Continuous vars

- For categorical X, NB gave us a linear discriminant
- What about continuous X?
  - e.g., Gaussian NB
- Will turn out the same, but we'll work it out for a generalization

$$\vec{X} \sim \underbrace{N(X \mid \mu, \Sigma)}_{\text{mean} \quad \text{covariance}} = 1/\sqrt{|2\pi\Sigma|}\, \exp\{-0.5\, \underbrace{(x-\mu)^T \Sigma^{-1} (x-\mu)}_{\text{quadratic}}\}$$

Multivariate Gaussians

svd



$\mu_1 = (0,3)$

$\Sigma_1 = U D_1 U^T$

$U = \begin{pmatrix} \sqrt{2}/2 & \frac{\sqrt{2}}{2} \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{pmatrix}$

$D_1 = \begin{pmatrix} .2 & 0 \\ 0 & .08 \end{pmatrix}$

$\sqrt{D} = \begin{pmatrix} .45 & 0 \\ 0 & .28 \end{pmatrix}$

$\Sigma_2 = U D_2 U^T$

$D_2 = \begin{pmatrix} .02 & 0 \\ 0 & .26 \end{pmatrix}$

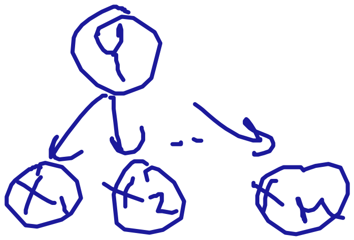
$\sqrt{D_2} = \begin{pmatrix} .14 & 0 \\ 0 & .51 \end{pmatrix}$

$\Sigma_3 = \begin{pmatrix} .1 & 0 \\ 0 & .1 \end{pmatrix} = D_3$

$\Sigma_3 = U D_3 U^T$

A generalization of Gaussian Naïve Bayes

$\Sigma_{ij} = \mathrm{cov}(X_i, X_j)$



GNB

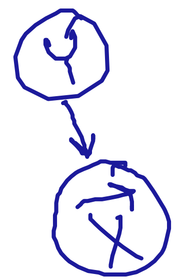
$X_j \sim N(\mu_j^Y, \sigma^2)$

$2M + 1$ parameters

(Y binary) (shared var)

or $4M + 1$ if no shared variances among classes, vars

$X_j \sim N(\mu_j^Y, (\sigma_j^Y)^2)$



$\vec{X} \sim N(\mu_Y, \Sigma)$

Σ: $M \times M$

μ_Y: μ_0 or μ_1 $(M \times 1)$

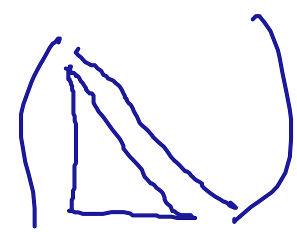
$2M + \binom{M}{2} + M$ params

$2M + 2\binom{M}{2} + 2M$ if no shared var among classes

Σ_0

Σ_1

make Σ diagonal



Generalizing GNB

$\log P(X \mid \mu, \Sigma)$

$-\log|2\pi\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)$

- $P(X \mid Y) = N(X \mid \mu_Y, \Sigma)$
  - if $\Sigma =$ $\sigma^2 I \Rightarrow$ GNB
- Pick $Y=1$ if
  - $P(Y=1)\, P(X \mid \mu_1, \Sigma) \geq P(Y=0)\, P(X \mid \mu_0, \Sigma)$

$$\log P(Y=1) - \log P(Y=0) + \underbrace{\log P(X \mid \mu_1, \Sigma) - \log P(X \mid \mu_0, \Sigma)} \geq 0$$

$$-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)\Sigma^{-1}(x-\mu_0)$$

$$\frac{1}{2}\left[-\cancel{x^T \Sigma^{-1} x} + 2x^T \Sigma^{-1}\mu_1 - \mu_1^T \Sigma \mu_1 + \cancel{x^T \Sigma^{-1} x} - 2x^T \Sigma^{-1}\mu_0 + \mu_0^T \Sigma^{-1}\mu_0 \right]$$

$$w_0 + w^T x \geq 0$$

Fisher linear discriminant

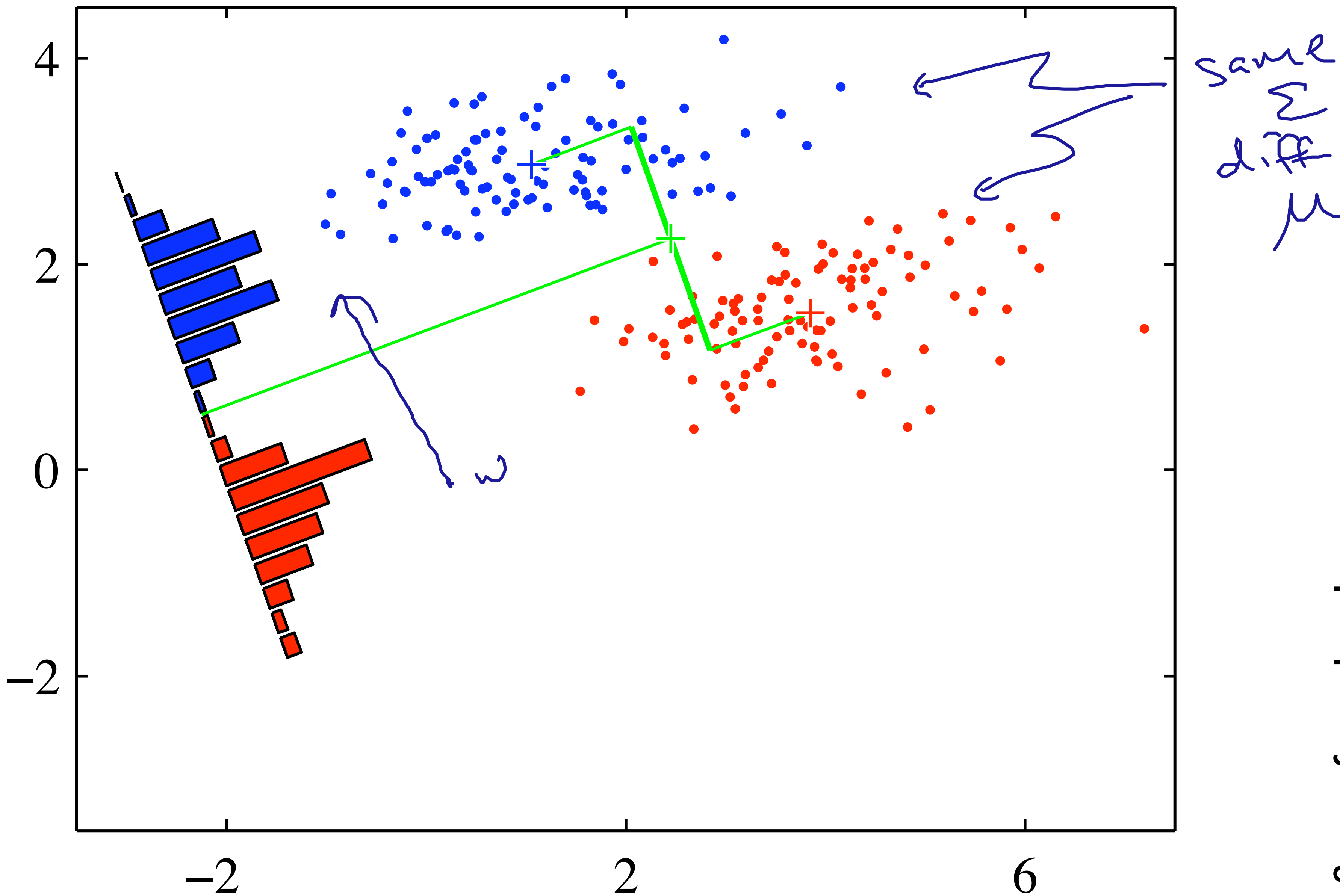


figure from book