

Machine Learning

4771

Instructor: Tony Jebara

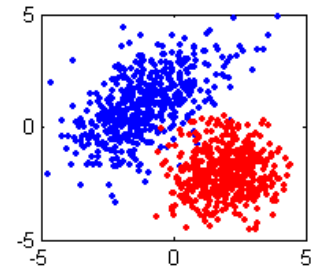
Topic 10

- Classification with Gaussians
- Regression with Gaussians
- Principal Components Analysis

Classification with Gaussians

- Have two classes, each with their own Gaussian:

$$\left\{ (x_1, y_1), \dots, (x_N, y_N) \right\} \quad x \in R^D \quad y \in \{0, 1\}$$



- Given parameters $\theta = \{\alpha, \mu_0, \Sigma_0, \mu_1, \Sigma_1\}$ we can generate iid data from $p(x, y | \theta) = p(y | \theta) p(x | y, \theta)$ by:

1) flipping a coin to get y via Bernoulli $p(y | \theta) = \alpha^y (1 - \alpha)^{1-y}$

2) sampling an x from y 'th Gaussian $p(x | y, \theta) = N(x | \mu_y, \Sigma_y)$

- Or, recover parameters from data using maximum likelihood

$$\begin{aligned} l(\theta) &= \log p(\text{data} | \theta) = \sum_{i=1}^N \log p(x_i, y_i | \theta) \\ &= \sum_{i=1}^N \log p(y_i | \theta) + \sum_{i=1}^N \log p(x_i | y_i, \theta) \\ &= \sum_{i=1}^N \log p(y_i | \alpha) + \sum_{y_i \in 0} \log p(x_i | \mu_0, \Sigma_0) + \sum_{y_i \in 1} \log p(x_i | \mu_1, \Sigma_1) \end{aligned}$$

Classification with Gaussians

- Max Likelihood can be done separately for the 3 terms

$$l = \sum_{i=1}^N \log p(y_i | \alpha) + \sum_{y_i \in 0} \log p(x_i | \mu_0, \Sigma_0) + \sum_{y_i \in 1} \log p(x_i | \mu_1, \Sigma_1)$$

- Count # of pos & neg examples (class prior): $\alpha = \frac{N_1}{N_0 + N_1}$
- Get mean & cov of negatives and mean & cov of positives:

$$\begin{aligned} \mu_0 &= \frac{1}{N_0} \sum_{y_i \in 0} x_i & \Sigma_0 &= \frac{1}{N_0} \sum_{y_i \in 0} (x_i - \mu_0)(x_i - \mu_0)^T \\ \mu_1 &= \frac{1}{N_1} \sum_{y_i \in 1} x_i & \Sigma_1 &= \frac{1}{N_1} \sum_{y_i \in 1} (x_i - \mu_1)(x_i - \mu_1)^T \end{aligned}$$

- Given (x,y) pair, can now compute likelihood $p(x, y)$
- To make classification, a bit of Decision Theory
- Without x, can compute prior guess for y $p(y)$
- Give me x, want y, I need posterior $p(y | x)$
- Bayes Optimal Decision: $\hat{y} = \arg \max_{y \in \{0,1\}} p(y | x)$
- Optimal iff we have true probability

Posterior gives Logistic

- Bayes Optimal Decision: $\hat{y} = \arg \max_{y=\{0,1\}} p(y | x)$

- To get conditional:

$$p(y | x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\sum_y p(x, y)} = \frac{p(x, y)}{p(x, y=0) + p(x, y=1)}$$

- Check which is greater: $p(y=0 | x) \geq ? \leq p(y=1 | x)$

- Or check if this is > 0.5 $p(y=1 | x) = \frac{p(x, y=1)}{p(x, y=0) + p(x, y=1)}$

$$= \frac{1}{\frac{p(x, y=0)}{p(x, y=1)} + 1}$$

$$= \frac{1}{\exp\left(-\log \frac{p(x, y=1)}{p(x, y=0)}\right) + 1}$$

- Get logistic squashing function of log-ratio of probability models

$$= \text{sigmoid}\left(\log \frac{p(x, y=1)}{p(x, y=0)}\right)$$

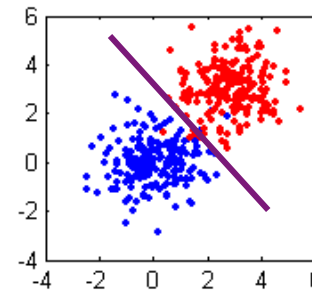
Linear or Quadratic Decisions

- Example cases, plotting decision boundary when $\alpha = 0.5$

$$\begin{aligned} p(y = 1 | x) &= \frac{p(x, y = 1)}{p(x, y = 0) + p(x, y = 1)} \\ &= \frac{\alpha N(x | \mu_1, \Sigma_1)}{(1 - \alpha) N(x | \mu_0, \Sigma_0) + \alpha N(x | \mu_1, \Sigma_1)} \end{aligned}$$

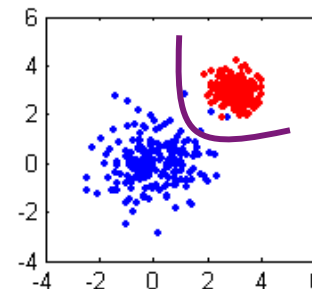
- If covariances are equal:

linear decision



- If covariances are different:

quadratic decision

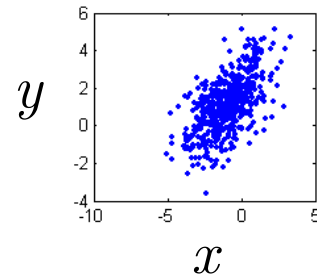


Regression with Gaussians

- Have input and output, each Gaussian:

$$\left\{ (x_1, y_1), \dots, (x_N, y_N) \right\} \quad x \in \mathbb{R}^{D_x} \quad y \in \mathbb{R}^{D_y}$$

concatenate $z_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$



$$p(z \mid \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} \sqrt{|\Sigma|}} \exp \left(-\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu) \right)$$

- Maximum Likelihood is as usual for a multivariate Gaussian

$$\mu = \frac{1}{N} \sum_{i=1}^N z_i \quad \Sigma = \frac{1}{N} \sum_{i=1}^N (z_i - \mu)(z_i - \mu)^T$$

- Bayes optimal decision:

$$\hat{y} = \arg \max_{y \in \mathbb{R}} p(y \mid x)$$

- Or we can use:

$$\hat{y} = E_{p(y|x)} \{y\}$$

- Have joint, need conditional:

$$p(y \mid x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\int_y p(x, y)}$$

Gaussian Marginals/Conditionals

•Conditional & marginal from joint: $p(y | x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\int_y p(x, y)}$

•Conditioning the Gaussian:

$$p(z | \mu, \Sigma) = \frac{1}{(2\pi)^{D/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(z - \mu)^T \Sigma^{-1}(z - \mu)\right)$$

$$p(x, y) = \frac{1}{(2\pi)^{D/2} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}\right)^T \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}\right)\right)$$

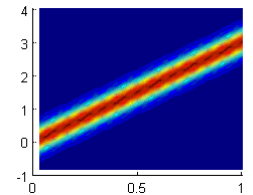
$$p(x) = \frac{1}{(2\pi)^{D_x/2} \sqrt{|\Sigma_{xx}|}} \exp\left(-\frac{1}{2}(x - \mu_x)^T \Sigma_{xx}^{-1}(x - \mu_x)\right)$$

$$= N(x | \mu_x, \Sigma_{xx})$$

$$p(y | x) = N\left(y | \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1}(x - \mu_x), \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}\right)$$

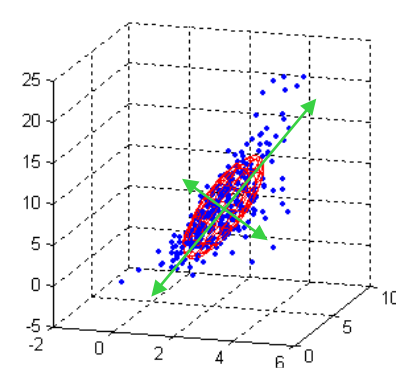
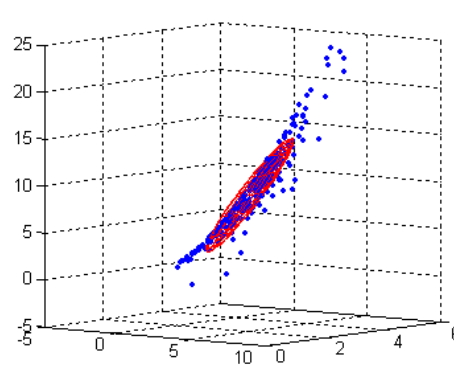
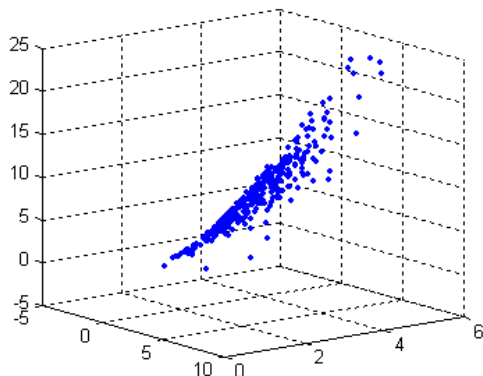
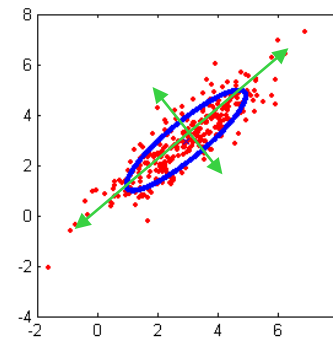
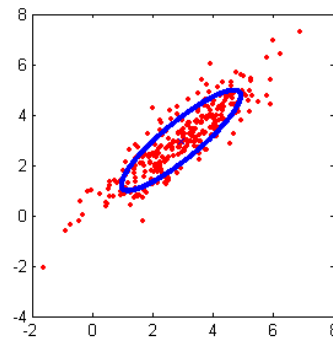
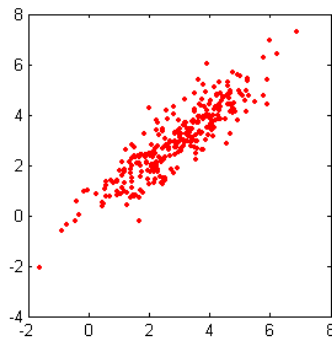
•Here argmax is expectation
which is conditional mean:

$$\hat{y} = \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1}(x - \mu_x)$$



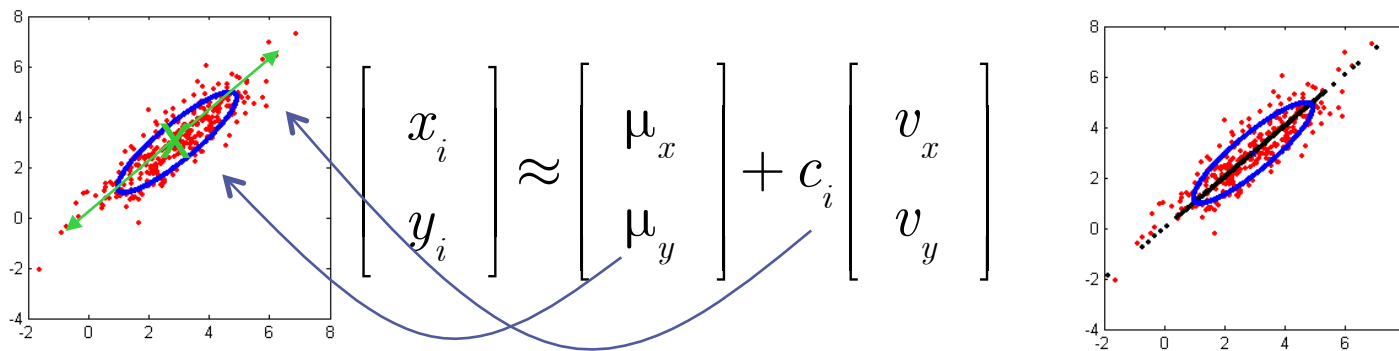
Principal Components Analysis

- Gaussians: for Classification, Regression... & Compression!
- Data can be constant in some directions, changes in others
- Use Gaussian to find directions of high/low variance



Principal Components Analysis

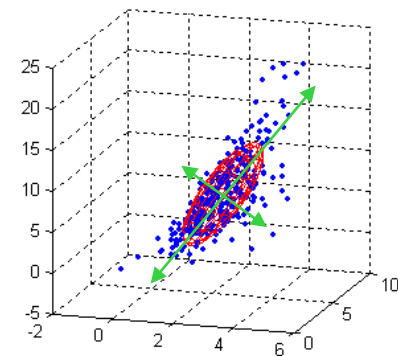
- Idea: instead of writing data in all its dimensions, only write it as mean + steps along one direction



- More generally, keep a subset of dimensions C from D (i.e. 2 of 3)

$$\vec{x}_i \approx \vec{\mu} + \sum_{j=1}^C c_{ij} \vec{v}_j$$

- Compression method: $\vec{x}_i \gg \vec{c}_i$
- Optimal directions: along eigenvectors of covariance
- Which directions to keep: highest eigenvalues (variances)



Principal Components Analysis

- If we have eigenvectors, mean and coefficients:

$$\vec{x}_i \approx \vec{\mu} + \sum_{j=1}^C c_{ij} \vec{v}_j$$

- Get eigenvectors (use eig() in Matlab): $\Sigma = V \Lambda V^T$

$$\begin{bmatrix} \Sigma(1,1) & \Sigma(1,2) & \Sigma(1,3) \\ \Sigma(1,2) & \Sigma(2,2) & \Sigma(2,3) \\ \Sigma(1,3) & \Sigma(2,3) & \Sigma(3,3) \end{bmatrix} = \begin{bmatrix} [\vec{v}_1] & [\vec{v}_2] & [\vec{v}_3] \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \begin{bmatrix} [\vec{v}_1] & [\vec{v}_2] & [\vec{v}_3] \end{bmatrix}^T$$

- Eigenvectors are orthonormal: $\vec{v}_i^T \vec{v}_j = \delta_{ij}$
- In coordinates of v , Gaussian is diagonal, $\text{cov} = \Lambda$
- All eigenvalues are non-negative $\lambda_i \geq 0$
- Higher eigenvalues are higher variance, use the top C ones

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \geq \dots$$

- To compute the coefficients: $c_{ij} = (\vec{x}_i - \vec{\mu})^T \vec{v}_j$

Eigenfaces

