

CS489/698

Lecture 6: Jan 23, 2017

Classification with Mixture of
Gaussians

[B] Sections 4.2, [M] Section 4.2

Linear Models

- Probabilistic Generative Models

Regression

Classification

Probabilistic Generative Model

- $\Pr(C)$: prior probability of class C
- $\Pr(\mathbf{x}|C)$: class conditional distribution of \mathbf{x}
- Classification: compute posterior $\Pr(C|\mathbf{x})$ according to Bayes' theorem

$$\begin{aligned}\Pr(C|\mathbf{x}) &= \frac{\Pr(\mathbf{x}|C) \Pr(C)}{\sum_C \Pr(\mathbf{x}|C) \Pr(C)} \\ &= k \Pr(\mathbf{x}|C) \Pr(C)\end{aligned}$$

Assumptions

- In classification, the number of classes is finite, so a natural prior $\Pr(C)$ is the multinomial

$$\Pr(C = c_k) = \pi_k$$

- When $\mathbf{x} \in \mathbb{R}^d$, then it is often OK to assume that $\Pr(\mathbf{x}|C)$ is Gaussian.
- Furthermore, assume that the same covariance matrix Σ is used for each class.

$$\Pr(\mathbf{x}|c_k) \propto e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}$$

Posterior Distribution

$$\begin{aligned} \Pr(c_k | \mathbf{x}) &= \frac{\pi_k e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}}{\sum_k \pi_k e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)}} \\ &= \frac{\pi_k e^{-\frac{1}{2}(\cancel{\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}} - 2\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k)}}{\sum_k \pi_k e^{-\frac{1}{2}(\cancel{\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}} - 2\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{x} + \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k)}} \end{aligned}$$

Consider two classes c_k and c_j

$$= \frac{1}{1 + \frac{\pi_j e^{\boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j}}{\pi_k e^{\boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k}}}}$$

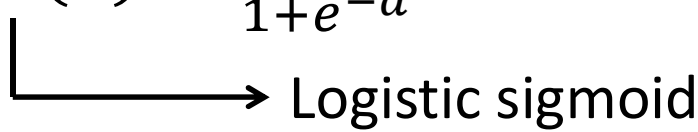
Posterior Distribution

$$\begin{aligned} &= \frac{1}{1+e^{-\left(\boldsymbol{\mu}_k^T - \boldsymbol{\mu}_j^T\right) \boldsymbol{\Sigma}^{-1} x + \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \ln \frac{\pi_k}{\pi_j}}} \\ &= \frac{1}{1+e^{-\left(\mathbf{w}^T x + w_0\right)}} \end{aligned}$$

where $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)$

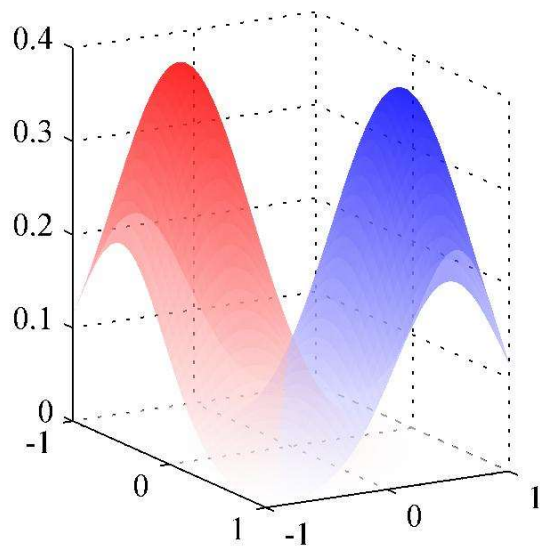
and $w_0 = -\frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \ln \frac{\pi_k}{\pi_j}$

Logistic Sigmoid

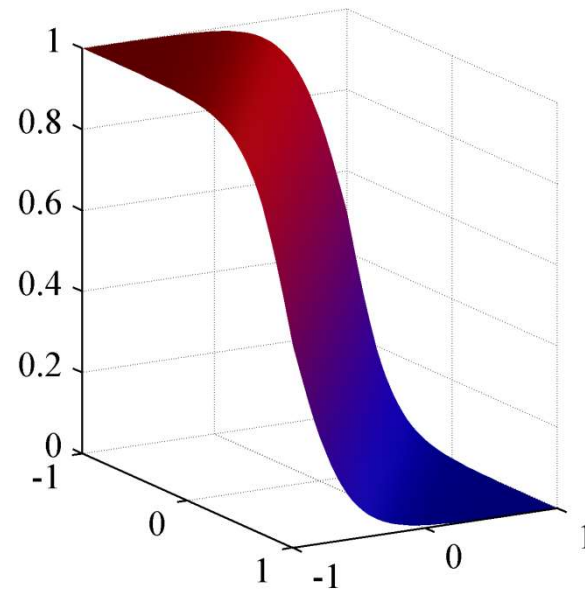
- Let $\sigma(a) = \frac{1}{1+e^{-a}}$
 Logistic sigmoid
- Then $\Pr(c_k|\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$
- Picture:

Logistic Sigmoid

class conditionals



posterior



Prediction

$$\begin{aligned} \text{best class} &= \operatorname{argmax}_k \Pr(c_k | \mathbf{x}) \\ &= \begin{cases} c_1 & \sigma(\mathbf{w}^T \mathbf{x} + w_0) \geq 0.5 \\ c_2 & \text{otherwise} \end{cases} \end{aligned}$$

Class boundary: $\sigma(\mathbf{w}_k^T \bar{\mathbf{x}}) = 0.5$

$$\Rightarrow \frac{1}{1 + e^{-(\mathbf{w}_k^T \bar{\mathbf{x}})}} = 0.5$$

$$\Rightarrow \mathbf{w}_k^T \bar{\mathbf{x}} = 0$$

\therefore linear separator

Multi-class problems

- When there are several classes, then posterior is a **softmax** (generalization of the sigmoid)

- Softmax: $\Pr(c_k | \mathbf{x}) = \frac{e^{f_k(\mathbf{x})}}{\sum_j e^{f_j(\mathbf{x})}}$

- Consider Gaussian conditional distributions

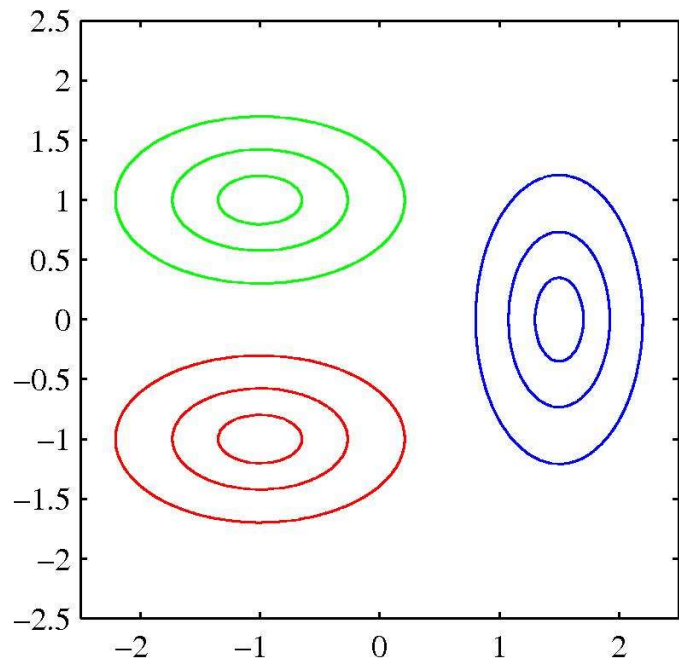
$$\Pr(c_k | \mathbf{x}) = \frac{\Pr(c_k) \Pr(\mathbf{x} | c_k)}{\sum_j \Pr(c_j) \Pr(\mathbf{x} | c_j)}$$

Softmax

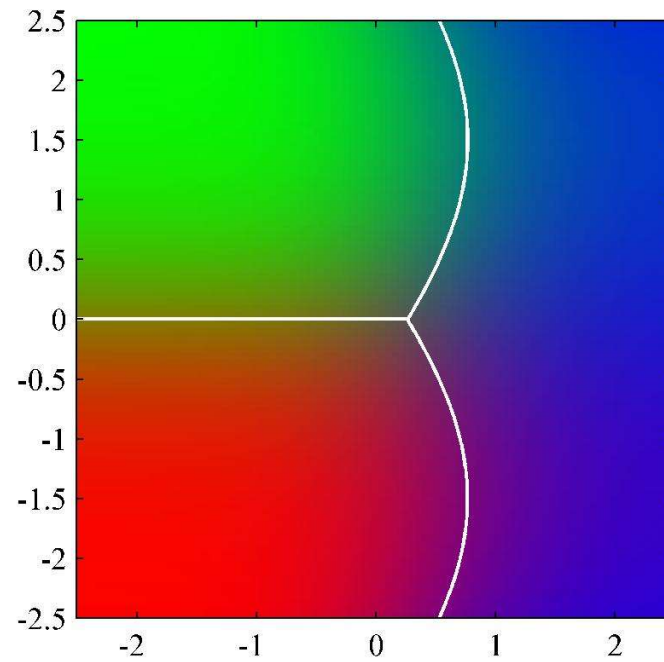
$$\begin{aligned} &= \frac{\pi_k e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}}{\sum_j \pi_j e^{-\frac{1}{2}(x-\mu_j)^T \Sigma^{-1}(x-\mu_j)}} \\ &= \frac{\pi_k e^{-\frac{1}{2}(-2\mu_k^T \Sigma^{-1}x + \mu_k^T \Sigma^{-1}\mu_k)}}{\sum_j \pi_j e^{-\frac{1}{2}(-2\mu_j^T \Sigma^{-1}x + \mu_j^T \Sigma^{-1}\mu_j)}} \\ &= \frac{e^{\mu_k^T \Sigma^{-1}x - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + \ln \pi_k}}{\sum_j e^{\mu_j^T \Sigma^{-1}x - \frac{1}{2}\mu_j^T \Sigma^{-1}\mu_j + \ln \pi_j}} \\ &= \frac{e^{w_k^T \bar{x}}}{\sum_j e^{w_j^T \bar{x}}} \implies \text{softmax} \end{aligned}$$

Softmax

class conditionals



posterior



Parameter Estimation

- Where do $\Pr(c_k)$ and $\Pr(\mathbf{x}|c_k)$ come from?
- Parameters: $\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$
 - $\Pr(c_1) = \pi, \quad \Pr(\mathbf{x}|c_1) \propto e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}$
 - $\Pr(c_2) = 1 - \pi, \quad \Pr(\mathbf{x}|c_2) \propto e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}$
- Estimate parameters by
 - **Maximum likelihood**
 - Maximum a posteriori
 - Bayesian learning

Maximum Likelihood Solution

- Likelihood:

$$L(\mathbf{X}, \mathbf{y}) = \Pr(\mathbf{X}, \mathbf{y} | \pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_n [\pi N(\mathbf{x}_n | \boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{y_n} [(1 - \pi) N(\mathbf{x}_n | \boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1 - y_n}$$

$y_n \in \{0, 1\}$

- ML hypothesis:

$$\langle \pi^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \boldsymbol{\Sigma}^* \rangle =$$

$$\operatorname{argmax}_{\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}} \sum_n y_n \left[\ln \pi - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_1) \right] \\ + (1 - y_n) \left[\ln(1 - \pi) - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_2) \right]$$

Maximum Likelihood Solution

- Set derivative to 0

$$0 = \frac{\partial \ln L(\mathbf{X}, \mathbf{y})}{\partial \pi}$$

$$\Rightarrow 0 = \sum_n y_n \left[\frac{1}{\pi} \right] + (1 - y_n) \left[-\frac{1}{1-\pi} \right]$$

$$\Rightarrow 0 = \sum_n y_n (1 - \pi) + (1 - y_n)(-\pi)$$

$$\Rightarrow \sum_n y_n = \pi (\sum_n y_n + \sum_n (1 - y_n))$$

$$\Rightarrow \sum_n y_n = \pi N$$

$$\boxed{\therefore \frac{\sum_n y_n}{N} = \pi}$$

Maximum Likelihood Solution

$$0 = \partial \ln L(\mathbf{X}, \mathbf{y}) / \partial \boldsymbol{\mu}_1$$

$$\Rightarrow 0 = \sum_n y_n [-\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_1)]$$

$$\Rightarrow \sum_n y_n \mathbf{x}_n = \sum_n y_n \boldsymbol{\mu}_1$$

$$\Rightarrow \sum_n y_n \mathbf{x}_n = N_1 \boldsymbol{\mu}_1$$

$$\therefore \frac{\sum_n y_n \mathbf{x}_n}{N_1} = \boldsymbol{\mu}_1$$

Similarly: $\frac{\sum_n (1-y_n) \mathbf{x}_n}{N_2} = \boldsymbol{\mu}_2$

Maximum Likelihood

$$\frac{\partial \ln L(\mathbf{X}, \mathbf{y})}{\partial \Sigma} = 0$$

$\Rightarrow \dots$

$$\Rightarrow \boxed{\Sigma = \frac{N_1}{N} \mathbf{S}_1 + \frac{N_2}{N} \mathbf{S}_2}$$

where $\mathbf{S}_1 = \frac{1}{N_1} \sum_{n \in c_1} (\mathbf{x}_n - \boldsymbol{\mu}_1)(\mathbf{x}_n - \boldsymbol{\mu}_1)^T$

$$\mathbf{S}_2 = \frac{1}{N_2} \sum_{n \in c_2} (\mathbf{x}_n - \boldsymbol{\mu}_2)(\mathbf{x}_n - \boldsymbol{\mu}_2)^T$$