Lecture 6: Classification with Mixtures of Gaussians
CS480/680 Intro to Machine Learning

2023-1-26

Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

Linear Models

- Probabilistic Generative Models

Regression | Classification

Probabilistic Generative Model

- $\Pr(C)$: prior probability of class C
- $\Pr(\boldsymbol{x}|C)$: class conditional distribution of $\boldsymbol{x}$

- Classification: compute posterior $\Pr(C|\boldsymbol{x})$ according to Bayes' theorem

$$\Pr(C|\boldsymbol{x}) = \frac{\Pr(\boldsymbol{x}|C)\Pr(C)}{\sum_C \Pr(\boldsymbol{x}|C)\Pr(C)}$$

$$= k Pr(\boldsymbol{x}|C)\Pr(C)$$

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

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

Posterior Distribution

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

$$= \frac{\pi_k e^{-\frac{1}{2}\left(\cancel{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}-2\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\right)}}{\sum_k \pi_k e^{-\frac{1}{2}\left(\cancel{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}-2\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_k\right)}}$$

Consider two classes c_k and c_j

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

Posterior Distribution

$$= \frac{1}{1+e^{-\left(\boldsymbol{\mu}_k^T-\boldsymbol{\mu}_j^T\right)\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k-\frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j-\ln\frac{\pi_k}{\pi_j}}}$$

$$= \frac{1}{1+e^{-\left(\boldsymbol{w}^T\boldsymbol{x}+w_0\right)}}$$

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

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

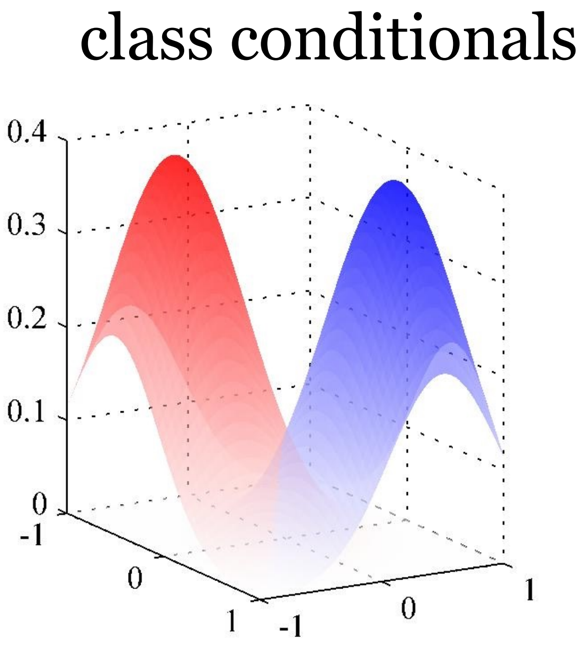
Logistic Sigmoid

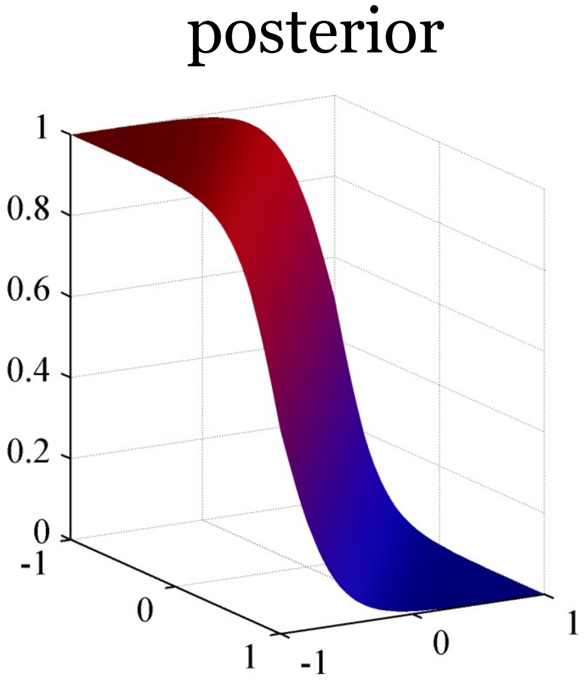
- Let $\sigma(a) = \frac{1}{1+e^{-a}}$

  → Logistic sigmoid

- Then $\Pr(c_k|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{x} + w_0)$
- Picture:

Logistic Sigmoid





Prediction

$$best\ class = argmax_k \Pr(c_k|\boldsymbol{x})$$

$$= \begin{cases} c_1 & \sigma(\boldsymbol{w}^T\boldsymbol{x} + w_0) \geq 0.5 \\ c_2 & \text{otherwise} \end{cases}$$

Class boundary: $\sigma\left(\boldsymbol{w}_k^T\overline{\boldsymbol{x}}\right) = 0.5$

$$\Longrightarrow \frac{1}{1+e^{-\left(\boldsymbol{w}_k^T\overline{\boldsymbol{x}}\right)}} = 0.5$$

$$\Longrightarrow \boldsymbol{w}_k^T\overline{\boldsymbol{x}} = 0$$

$\therefore$ **linear separator**

Multi-class Problems

- Consider Gaussian conditional distributions with identical Σ

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

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

$$= \frac{\pi_k e^{-\frac{1}{2}\left(-2\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k\right)}}{\sum_j \pi_j e^{-\frac{1}{2}\left(-2\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j\right)}}$$

$$= \frac{e^{\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{u}_k+\ln\pi_k}}{\sum_j e^{\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j+\ln\pi_j}} = \frac{e^{\boldsymbol{w}_k^T\bar{\boldsymbol{x}}}}{\sum_j e^{\boldsymbol{w}_j^T\bar{\boldsymbol{x}}}} \implies \text{softmax}$$

where $\boldsymbol{w}_k^T = (-\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln\pi_k,\ \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1})$

Softmax

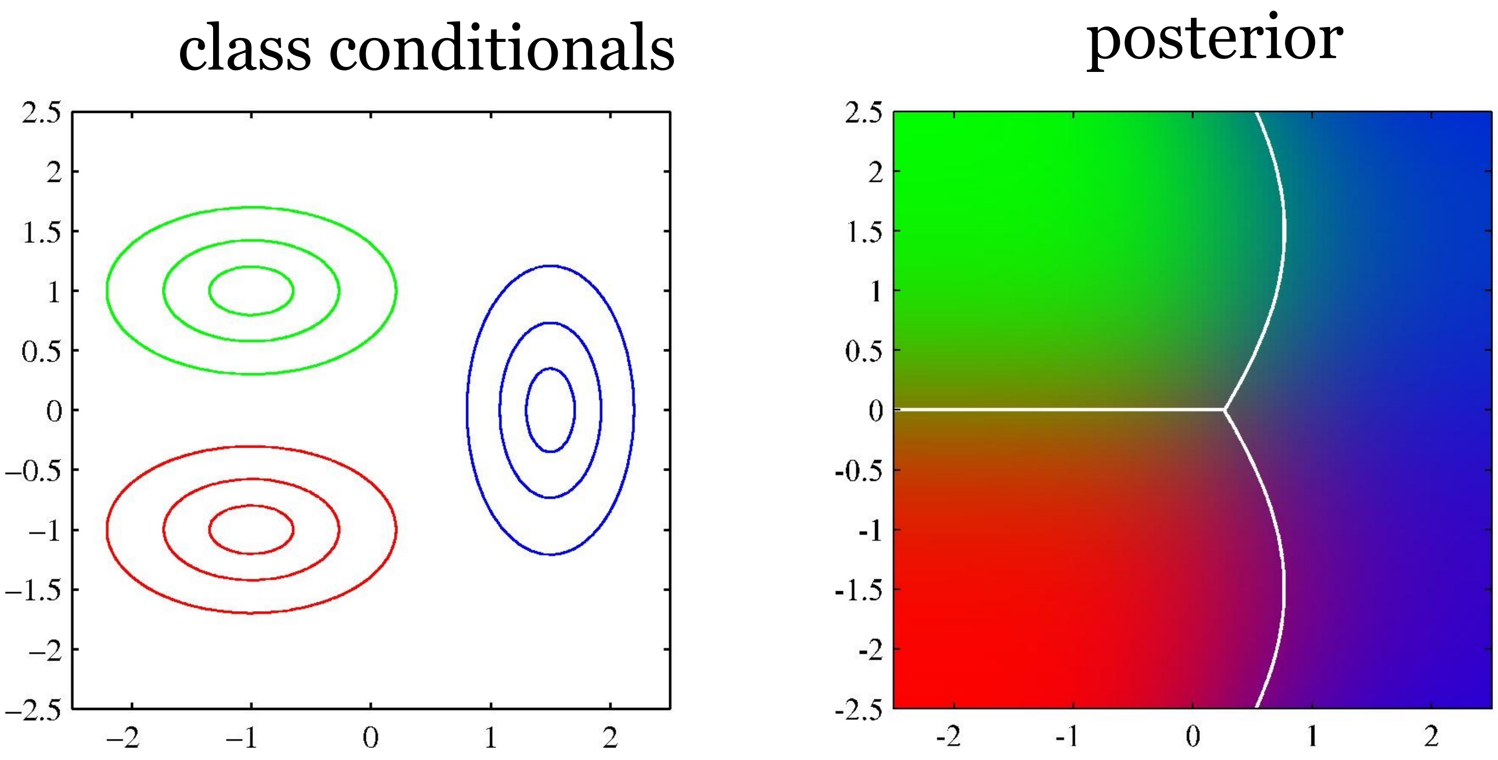
- When there are several classes, the posterior is a **softmax** (sigmoid generalization)
- Softmax distribution: $\Pr(c_k|\boldsymbol{x}) = \frac{e^{f_k(x)}}{\sum_j e^{f_j(x)}}$
- Argmax distribution:

$$\Pr(c_k|\boldsymbol{x}) = \begin{cases} 1 & \text{if } k = argmax_j\, f_j(x) \\ 0 & \text{otherwise} \end{cases}$$

$$= \lim_{base \to \infty} \frac{base^{f_k(x)}}{\sum_j base^{f_j(x)}}$$

$$\approx \frac{e^{f_k(x)}}{\sum_j e^{f_j(x)}} \quad \text{(softmax approximation)}$$

Softmax


class conditionals
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
−1
0
1
2
posterior
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1
0
1
2

Parameter Estimation

- Where do $\Pr(c_k)$ and $\Pr(\boldsymbol{x}|c_k)$ come from?

- Parameters: π, $\boldsymbol{\mu_1}$, $\boldsymbol{\mu_2}$, $\boldsymbol{\Sigma}$

  $\Pr(c_1) = \pi$, $\quad \Pr(\boldsymbol{x}|c_1) = k_{\boldsymbol{\Sigma}}\, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_1})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)}$

  $\Pr(c_2) = 1 - \pi$, $\quad \Pr(\boldsymbol{x}|c_2) = k_{\boldsymbol{\Sigma}}\, e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_2})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)}$

  where $k_{\boldsymbol{\Sigma}}$ is the normalization constant that depends on $\boldsymbol{\Sigma}$

- Estimate parameters by
  - **Maximum likelihood**
  - Maximum a posteriori
  - Bayesian learning

Maximum Likelihood Solution

- Likelihood:

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

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

- ML hypothesis:

$$< \pi^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \boldsymbol{\Sigma}^* > = argmax_{\pi, \boldsymbol{\mu}_2, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}} \sum_n y_n \left[\ln \pi + \ln k_{\boldsymbol{\Sigma}} - \frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_1) \right]$$

$$+ (1 - y_n) \left[\ln(1-\pi) + \ln k_{\boldsymbol{\Sigma}} - \frac{1}{2} (\boldsymbol{x}_n - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_2) \right]$$

Maximum Likelihood Solution

- Set derivative to 0

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

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

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

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

$$\Longrightarrow \sum_n y_n = \pi N \quad \text{(where } N \text{ is the \# of training points)}$$

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

UNIVERSITY OF WATERLOO

Maximum Likelihood Solution

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

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

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

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

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

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

where N_1 is the # of data points in class 1

N_2 is the # of data points in class 2

Maximum Likelihood

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

$$\Longrightarrow \cdots$$

$$\Longrightarrow \boxed{\Sigma = \frac{N_1}{N}\boldsymbol{S}_1 + \frac{N_2}{N}\boldsymbol{S}_2}$$

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

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

($\boldsymbol{S}_k$ is the empirical covariance matrix of class k)