Li Yu

Hongda Mao

Joan Wang

10.5 Unsupervised Bayesian Learning

The Bayes Classifier

- Maximum-likelihood methods:

  parameter vector θ is a fixed but unknown value

- Bayes methods:

  parameter vector θ is a random variable with known prior distribution

The Bayes Classifier

Bayes Formula:

information from samples

$$P(w_k \mid x, D) = \frac{P(x \mid w_k, D) P(w_k \mid D)}{\sum_{i=1}^{c} P(x \mid w_i, D) P(w_i \mid D)} = \frac{P(x \mid w_k, D) P(w_k)}{\sum_{i=1}^{c} P(x \mid w_i, D) P(w_i)}$$

Assumptions:

- 1. Classes number c is known.
- 2. Prior probabilities $P(w_k)$ for each class are known, k = 1, . . . , c.
- 3. Forms of the class-conditional probability densities $P(x \mid w_k, \theta_k)$ are known.
- 4. Part of our knowledge about θ is from a known prior density $P(\theta)$.
- 5. Rest knowledge about θ is from the set of samples D. $p(\mathbf{x} \mid \theta) = \sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j, \theta_j) P(\omega_j)$

The Bayes Classifier

$$P(w_k \mid x, D) = \frac{P(x \mid w_k, D) P(w_k)}{\sum_{i=1}^{c} P(x \mid w_i, D) P(w_i)}$$

unknown

- Based on the assumptions, we have

$$P(x \mid w_k, D) = \int P(x, \theta \mid w_k, D) d\theta$$
$$= \int P(x \mid \theta, w_k, D) P(\theta \mid w_k, D) d\theta = \int P(x \mid w_k, \theta_k) P(\theta \mid D) d\theta$$

where $P(x \mid \theta, w_k, D) = P(x \mid w_k, \theta_k)$ and $P(\theta \mid w_k, D) = P(\theta \mid D)$

x is independent of the samples

class has nothing to do with distribution of θ

The Bayes Classifier

$$P(x \mid w_k, D) = \int P(x \mid w_k, \theta_k) P(\theta \mid D) d\theta$$

- Estimate of $P(x \mid w_k)$ is obtained by averaging $P(x \mid w_k, \theta_k)$ over θ_k
- The task at hand now is to estimate $P(\theta \mid D)$ from the sample set D.

The basic equations for unsupervised Bayesian learning

- The Posterior Density :

$$p(\theta \mid D) = \frac{p(D \mid \theta) p(\theta)}{\int p(D \mid \theta) p(\theta) d\theta}$$

- The likelihood yielded by the samples:

$$p(D \mid \theta) = \prod_{k=1}^{n} p(x_k \mid \theta)$$

- The Posterior Density in recursive form:

$$p(\theta \mid D^n) = \frac{p(x_n \mid \theta) p(\theta \mid D^{n-1})}{\int p(x_n \mid \theta) p(\theta \mid D^{n-1}) d\theta}$$

The relation between Bayesian and the M-L solutions

- Again, the posterior density:

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int p(D \mid \theta)p(\theta)d\theta}$$

- This equation emphasizes the relation between the Bayesian and the maximum-likelihood solutions.

If $P(\theta)$ is essentially uniform over the region where $P(D \mid \theta)$ peaks, then $p(\theta \mid D)$ peaks at the same place.

➢1. The only significant peak occurs at $\theta = \hat{\theta}$

➢2. The peak is very sharp

- We can get:

$$P(x \mid w_i, D) = \int P(x \mid w_i, \theta_i)P(\theta \mid D)d\theta$$

$$P(w_i \mid x, D) = \frac{P(x \mid w_i, D)P(w_i)}{\sum_{j=1}^{c} P(x \mid w_j, D)P(w_j)}$$

$$\Rightarrow$$

$$P(x \mid w_i, D) \approx P(x \mid w_i, \hat{\theta})$$

$$P(w_i \mid x, D) \approx \frac{P\left(x \mid w_i, \hat{\theta}\right)P(w_i)}{\sum_{j=1}^{c} P\left(x \mid w_j, \hat{\theta}\right)P(w_j)}$$

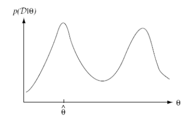
The relation between Bayesian and the M-L solutions cont'd

- Conclusions: The use of the maximum-likelihood estimate $\hat{\theta}$ as if it were the true value of θ in designing the Bayes classifier.

- Discussions:

➢If there are a large amounts of data, maximum-likelihood and Bayes methods will agree (or nearly agree).

➢If there are only a small amounts of data, there exist some small problems where the approximations are poor.



$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int p(D \mid \theta)p(\theta)d\theta}$$

Supervised and unsupervised learning

- Differences:

➢Lack of identifiability (main difference)

With supervised learning, it merely means that instead of obtaining a unique parameter vector we obtain an equivalence class of parameter vectors.

With unsupervised learning, it may cause serious problems. When θ can't be determined uniquely, the mixture can't be decomposed into its true components.

➢Computational complexity

Supervised learning: the possibility of finding sufficient statistics allows solutions that are analytically pleasing and computationally feasible.

Unsupervised learning: there is no way to avoid the fact that the samples are obtained from a mixture density.

$$p(\mathrm{x} \mid \theta) = \sum_{j=1}^{c} p(\mathrm{x} \mid w_j, \theta_j)P(w_j)$$

Supervised and unsupervised learning cont'd

$$p(\mathrm{x} \mid \theta) = \sum_{j=1}^{c} p(\mathrm{x} \mid w_j, \theta_j)P(w_j)$$

From the above equation, it is not easy for us to find a simple exact solutions for $p(\theta \mid D)$

Again, the likelihood

$$p(D \mid \theta) = \prod_{k=1}^{n}[\sum_{j=1}^{c} p(\mathrm{x}_k \mid w_j, \theta_j)p(w_j)]$$

1. $p(D \mid \theta)$ is the sum of c^n products of component densities. Each term in this sum can be interpreted as the joint probability of obtaining the samples $x_1, \ldots x_n$.
2. If the component densities do not overlap, thus as varies, only one term in the mixture density is nonzero.

Supervised and unsupervised learning cont'd

➢Another way to compare supervised and unsupervised learning :

$$p(\theta \mid D^n) = \frac{p(x_n \mid \theta)p(\theta \mid D^{n-1})}{\int p(x_n \mid \theta)p(\theta \mid D^{n-1})d\theta}$$

- Substitute the mixture density for $p(x_n \mid \theta)$

$$p(\theta \mid D^n) = \frac{\sum_{j=1}^{c} p(x_n \mid w_j, \theta_j)}{\sum_{j=1}^{c}\int p(x_n \mid w_j, \theta_j)p(\theta \mid D^{n-1})d\theta} p(\theta \mid D^{n-1})$$

Unsupervised

- Let $p(w_1)=1$ and all the other prior probabilities are zero.

$$p(\theta \mid D^n) = \frac{p(x_n \mid w_1, \theta_1)}{\int p(x_n \mid w_1, \theta_j)p(\theta \mid D^{n-1})d\theta} p(\theta \mid D^{n-1})$$

Supervised

Example 1: Unsupervised learning of Gaussian Data

- Consider the one-dimensional, two-component mixture with

$p(x \mid w_1) \sim N(\mu, 1)$ $\quad p(x \mid w_2, \theta) \sim \mathrm{N}(\theta, 1)$

where μ, $P(w_1)$ and $P(w_2)$ are known.

$$p(x \mid \theta) = \frac{P(w_1)}{\sqrt{2\pi}}\exp[-\frac{1}{2}(x-\mu)^2] + \frac{P(w_2)}{\sqrt{2\pi}}\exp[-\frac{1}{2}(x-\theta)^2]$$

We seek the mean of the second component. Suppose that the prior density $p(\theta)$ is uniform from a to b, Then after one observation ($x=x_1$) we can get.

$$p(\theta \mid x_1) = \alpha p(x_1 \mid \theta)p(\theta) = \begin{cases} \alpha'\{P(w_1)\exp[-\frac{1}{2}(x_1-\mu)^2] + \\ P(w_2)\exp[-\frac{1}{2}(x_1-\theta)^2]\} & \mathrm{a} \le \theta \le \mathrm{b} \\ 0 & \text{otherwise} \end{cases}$$

+ Example 1 cont'd

$$p(\theta \mid x_1) = \alpha p(x_1 \mid \theta) p(\theta) = \begin{cases} \alpha' \{ P(w_1) \exp[-\frac{1}{2}(x_1 - \mu)^2] + \\ P(w_2) \exp[-\frac{1}{2}(x_1 - \theta)^2] \} & a \le \theta \le b \\ 0 & \text{otherwise} \end{cases}$$

- Discussions:

$$\text{the place of peaks of } p(\theta \mid x_1) = \begin{cases} \theta = x_1 & a \le x_1 \le b \\ \theta = a & x_1 < a \\ \theta = b & x_2 > b \end{cases}$$

+ Example 1 cont'd

Consider a second sample x_2

$$P(\theta \mid x_1, x_2) = \beta p(x_2 \mid \theta) p(\theta \mid x_1)$$

$$= \begin{cases} \beta' \{ P(w_1)P(w_1) \exp[-\frac{1}{2}(x_1 - \mu)^2 - \frac{1}{2}(x_2 - \mu)^2] \\ + P(w_1)P(w_2) \exp[-\frac{1}{2}(x_1 - \mu)^2 - \frac{1}{2}(x_2 - \theta)^2] \\ + P(w_2)P(w_1) \exp[-\frac{1}{2}(x_1 - \theta)^2 - \frac{1}{2}(x_2 - \mu)^2] \\ + P(w_2)P(w_2) \exp[-\frac{1}{2}(x_1 - \theta)^2 - \frac{1}{2}(x_2 - \theta)^2] \\ \quad a \le \theta \le b \\ 0 \quad \text{otherwise} \end{cases}$$

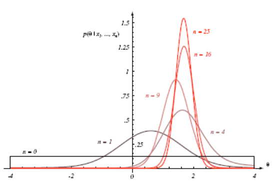
Unfortunately, the primary thing we learn from this expression is already

Complicated when n=2. With n samples there will be 2^n terms, as a result , the computational cost will be very heavy.

+ Example 1 cont'd

So it is possible to use the following relation.

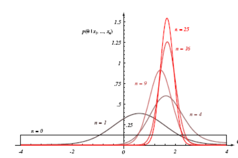
$$p(\theta \mid D^n) = \frac{p(x_n \mid \theta) p(\theta \mid D^{n-1})}{\int p(x_n \mid \theta) p(\theta \mid D^{n-1}) d\theta}$$

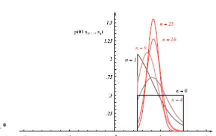


+ Example 1 cont'd

Discussion:

One of the main differences between the Bayesian and the maximum-likelihood approaches to unsupervised learning is the presence of the prior density $p(\theta)$





+ Decision-Directed Approximation

- **Why:** both maximum-likelihood and the bayesian methods have high computational requirements.
- **Solutions:** because the difference between supervised and unsupervised learning is the presence of labels, it is natural to propose the following:
  - Use prior information to train a classifier.
  - Label new data with this classifier.
  - Use the new labeled samples to train a new (supervised) classifier.
- This approach is known as the **decision-directed** approach [1] to unsupervised learning.
- Obvious **limitations** include:
  - If the initial classifier is not reasonably good, the process can diverge.
  - The tails of the distribution tend not to be modeled well this way, which results in significant overlap between the component densities.
- In practice, this approach works well because it is easy to leverage previous work for the initial classifier.
- Also, it is less computationally expensive than the pure Bayesian unsupervised learning approach.
- *[1] "Classical adaptive algorithms (LMS, RLS, CMA, decision directed) seen as recursive structures" by P. Duhamel, M. Montazeri and K. Hilal.*

+ 10.7 Criterion Functions For Clustering

+ Criterion Functions for Clustering

- Purpose: measures the clustering quality of any partition of the data.
- Suppose: a set D of n samples X1, X2,…Xn is classified into c clusters D1, D2,…Dc.
  - Samples in the same cluster are more similar than samples in different clusters.
  - Finds a the partition that optimizes the criterion function

+ The Sum-of-Squared-Error Criterion

- Simplest and most widely used one:

$m_i = \frac{1}{n_i}\sum_{x\in D_i} x$ where n_i is the number of samples in D_i

m_i is the mean of the samples

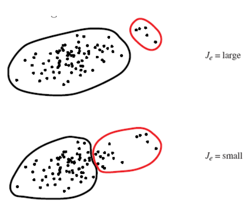
Sum-of-squared errors:

$J_e = \sum_{i=1}^{c}\sum_{x\in D_i} \| x - m_i \|^2$ c is the number of clusters

The optimal partitioning is defined as one that minimizes J_e (minimum variance)

+ The Sum-of-Squared-Error Criterion

- Problem:
  - Work well when clusters form compact clouds
  - Fail when there are great differences in the number of samples in different clusters.



+ Related Criteria

- Rewrite the criterion function

$$J_e = \sum_{i=1}^{c}\sum_{x\in D_i} \| x - m_i \|^2 \quad \Leftrightarrow \quad J_e = \frac{1}{2}\sum_{i=1}^{c} n_i \bar{s}_i$$

where

$$m_i = \frac{1}{n_i}\sum_{x\in D_i} x$$

where

$$\bar{s}_i = \frac{1}{n_i^2}\sum_{x\in D_i}\sum_{x'\in D_i} \| x - x' \|^2$$

S_i is the average squared distance between points in the i-*th* cluster (a similarity function)

+ Related Criteria - Continued

- The similarity function can be replaced by other appropriate similarity functions
- e.g. the average, the median or even the maximum distance between points in a cluster

+ 10.7.3 Scatter Criteria – Another type of Criterion Functions

- Definition of Scatter Matrix
  - Scatter matrix for the i-*th* cluster

$$S_i = \sum_{x\in D_i} (x - m_i)(x - m_i)^t$$

  - Within-cluster scatter matrix

$$S_W = \sum_{i=1}^{c} S_i$$

  - Between-cluster scatter matrix

$$S_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t$$

whereas

$$m_i = \frac{1}{n}\sum_{x\in D_i} x$$

(Cluster mean)

$$m = \frac{1}{n}\sum_{i=1}^{c} n_i m_i$$

(Total mean)

+

- Total Scatter Matrix

whereas

$$S_T = \sum_{x \in D} (x - m)(x - m)^t \qquad m = \frac{1}{n}\sum_{i=1}^{c} n_i m_i$$

(Total mean)

- Is the sum of within-cluster scatter matrix and between-cluster scatter matrix
- Not dependent on the formation of clusters
- Only dependent on all samples

+ How to measure a scatter matrix?

Trace Criterion	Determinant Criterion	Invariant Criterion
• Sum of diagonal elements	• Determinant of the matrix	•Appropriate functions of eigenvalues

+ Trace Criterion

- Minimize the sum of diagonal elements of S_w - $tr[S_W]$
- Or maximize $tr[S_B]$

Interesting fact:

$$tr[S_W] = \sum_{i=1}^{c} tr[S_i] = \sum_{i=1}^{c} \sum_{x \in D_i} \| x - m_i \|^2 = J_e$$

+ Determinant Criterion

- Minimize the determinant of S_W - $|S_w|$

$$J_d = |S_W| = \left| \sum_{i=1}^{c} S_i \right|$$

- S_B is not chosen because it will become singular if the number of clusters is less than or equal to the dimensionality
- Minimizing J_d is similar to minimizing J_e , but not necessarily the same

+ Invariant Criteria

- Maximize $tr[S_W^{-1} S_B] = \sum_{i=1}^{d} \lambda_i$
- Or minimize $J_f = tr[S_T^{-1} S_W] = \sum_{i=1}^{d} \frac{1}{1+\lambda_i}$

Whereas λ_i are eigenvalues of $S_W^{-1} S_B$

Eigenvalue – a scalar value for linear transformation that only changes the eigenvector's length but not direction

+

Trace Criterion	Determinant Criterion	Invariant Criterion
• Sum of diagonal elements	• Determinant of the matrix	• Appropriate functions of eigenvalues

Are they the same?

Trace criterion
(Tend to equal the number of points in each cluster)

Determinant criterion
(Tend to form large and small clusters)

Invariant criterion
(subtle)

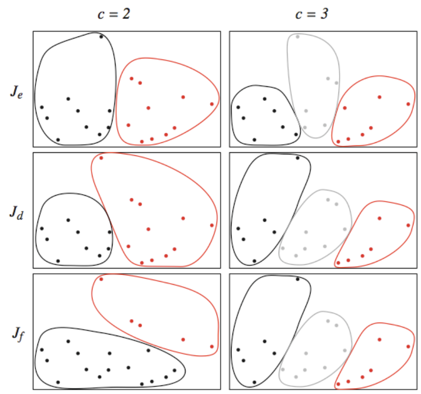


Figure from R. O. Duda, P. E. Hart, *Pattern classification*, 2001.

A Comparison of Cluster Validity Criteria For a Mixture of Normal Distributed Data [2]

- Clustering experiment based on 21 different criteria for simulated Gaussian data sets
- Conclusion: the most reliable criteria among the ones that they tested were:

(1) The trace average density criterion (trace of fuzzy covariance matrix)

(2) The Steinberg±Zeitouni criterion [3]

(3) The modified trace criterion ($tr[S_W]/c$).

[2] A. Geva et al, "A comparison of cluster validity criteria for a mixture of normal distributed data", *Pattern Recognition Letters*, 2000.
[3] Y. Steinberg and O. Zeitouni, "On tests for normality", *IEEE Trans. Inform. Theory*, Vol. 38, 1992.

Questions?

Thank you ☺