Expectation Maximization

7/7 得分（100%）

测验通过！

返回第 #4 周课程

---

1 / 1 分

1。
Bayesian Clustering using Normal Distributions.

Suppose we are doing Bayesian clustering with K classes, and multivariate normal distributions as our class-conditional distributions.

Let $\mathbf{X} \in \mathbf{R}^n$ represent a single data point, and $C \in \{1, 2, \ldots, K\}$ its unobserved class.

Which of the following statement(s) is/are always true in the general case?

- $P(\mathbf{X}|C = c) \sim N(0, I_n) \quad \forall c \in \{1, 2, \ldots, K\}$, where 0 is the all-zero vector and I_n is the $n \times n$ identity matrix.
- $P(\mathbf{X}|C = c) \sim N(\mu_c, I_n) \quad \forall c \in \{1, 2, \ldots, K\}$, for some class-specific parameter μ_c that represents the distribution of data coming from the class c. (I_n is the $n \times n$ identity matrix.)
- $P(\mathbf{X}|C = c) \sim N(\mu_c, \Sigma_c) \quad \forall c \in \{1, 2, \ldots, K\}$, for some class-specific parameters μ_c and Σ_c that represent the distribution of data coming from the class c.

正确

This is the definition of having a multivariate normal distribution as the class-conditional distribution: given the class from which the data point came from, the distribution of the data point follows a multivariate normal distribution with mean and covariance parameters specific to its particular class.

- $X_i \perp X_j \quad \forall i, j \in \{1, 2, \ldots, n\}, i \neq j$, i.e., for any given data point, knowing one coordinate does not give us any information about another coordinate.
- $X_i \perp X_j \mid C \quad \forall i, j \in \{1, 2, \ldots, n\}, i \neq j$, i.e., for any given data point,

  if we know which class the data point comes from, then knowing one coordinate does not give us any information about another coordinate.

---

1 / 1 分

2。

Hard Assignment EM. Continuing from the previous question, let us now fix each class-conditional distribution to have the identity matrix as its covariance matrix.

If we use hard-assignment EM to estimate the class-dependent mean vectors, which of the following can we say about the resulting algorithm?

- It is an an algorithm that cannot be viewed as an instance of k-means.
- It is equivalent to running standard k-means clustering with K clusters.

正确

You will always assign vertices to their closest (in Euclidean distance) cluster centroid, just as in k -means.

- It is an instance of k-means, but using a different

distance metric rather than standard Euclidean distance.

---

1 / 1 分

3。

***Hard Assignment EM.** Now suppose that we fix each class-conditional distribution to have the same diagonal matrix D as its covariance matrix, where D is **not** the identity matrix.

If we use hard-assignment EM to estimate the class-dependent mean vectors, which of the following can we say about the resulting algorithm?

- (●) It is an instance of k-means, but using a different distance metric rather than standard Euclidean distance.

正确

You will always assign vertices to their closest cluster centroid, just as in k -means. But here the definition of ``closest'' is skewed by the covariance matrix so that it does not equally depend on each dimension and is thus not a Euclidean distance.

- ( ) It is equivalent to running standard k-means clustering with K clusters.
- ( ) It is an an algorithm that cannot be viewed as an instance of k-means.

---

1 / 1 分

4。

EM Running Time. Assume that we are trying to estimate parameters for a Bayesian network structured as a binary (directed) tree (**not** a polytree) with n variables, where each node has at most one parent. We parameterize the network using table CPDs. Assume that each variable has d values. We have a data set with M instances, where some observations in each instances are missing. What is the tightest asymptotic bound you can give for the worst case running time of EM on this data set for K iterations? In this and following questions, concentrate on the EM part of the learning algorithm only. You don't need to consider the running time of additional steps if the full learning algorithm needs any.

- $O(KMn^2d)$
- $O(KMn^2d^2)$
- $O(KMnd^2)$

正确

At each iteration and for every instance, it is required to run exact inference over the given network. Using clique-tree calibration, the cost of inference is the number of cliques (n) times the size of the clique potential which is d^2 (due to the tree-structure of the network each clique can have only 2 variables in its scope).

- Can't tell using only the information given

---

1 / 1 分

5。

EM Running Time. Use the setting of the previous question, but now we assume that the network is a polytree, in which some variables have several parents. What is the cost of running EM on this data set for K iterations?

- $O(KMn^2d^2)$
- $O(KMn^2d)$
- $O(KMnd^2)$
- Can't tell using only the information given.

正确

We cannot tell because now the factors in the clique tree can be considerably larger than d^2 (but we do not how much larger they might be).

---

1 / 1 分

6。

***Optimizing EM.**

Now, going back to the tree setting, where each node has at most one parent, assume that we are in a situation where at most 2 variables in each data instance are unobserved (not necessarily the same 2 in each instance). Can we implement EM more efficiently? If so, which of the following reduced complexities can you achieve?

- No computational savings can be achieved.
- $O(KMd^2)$
- $O(K(M+n)d^2)$

正确

In this case, the cost of the E-step is Md^2 since we can easily compute the probabilities of each possible completion of instances when only up to 2 variables are missing. We can use this to compute the expected sufficient statistics (where we will be summing over the M instances and up to d^2 possible completed instances). The cost of the M-step will be nd^2 (it is equal to the number of parameter values computed).

- $O(KMnd^2)$

---

1 / 1 分

7。

***Optimizing EM.**

Still in the tree setting, now assume that we are in a situation where at most 2 variables in each data instance are unobserved, but it's the same 2 each instance. Can we implement EM more efficiently? If so, which of the following reduced complexities can you achieve?

- [x] $O(KMd^2)$

正确

In this case, most of the graph is conditionally independent of the unobserved variables, so we can restrict our EM process to the sub-graph consisting of the unobserved variables and their Markov blankets, and fix parameters for the rest of the network once at the beginning, using standard MLE. Thus, the cost of updating a small subset of the parameters at each M-step will be no more than $O(d^2)$. In the E-step, for each instance, we will run inference over a small subset of variables at a cost of d^2 per instance. Accordingly, the cost of the E-step will be Md^2.

- [ ] No computational savings can be achieved.
- [ ] $O(KMnd^2)$
- [ ] $O(K(M+n)d^2)$