ECE M146 Introduction to Machine Learning

Prof. Lara Dolecek

ECE Department, UCLA

Today's Lecture

Recap:

- Unsupervised Learning: K-means and PCA

New topic:

- Mixture modeling
- Expectation maximization; soft K-means

Today's Lecture

Recap:

- Unsupervised Learning: K-means and PCA

New topic:

- Mixture modeling
- Expectation maximization; soft K-means

Recap: K-means

- We have already seen two popular instances of methods for unsupervised learning.
- K-means is an iterative method for clustering
  - Key idea: identify cluster representatives (prototype) and assign each data point to the cluster of the closest prototype, so that the sum of all distances from data points to their prototypes is minimized.
  - How: iterate between finding the closest prototype based on the given prototypes, for each data point, and adjusting the prototype position based on the points assigned to its cluster.

Recap: PCA

- We have already seen two popular instances of methods for unsupervised learning.
- PCA is a projection method for dimensionality reduction
  - Key idea: project onto the dimension of the highest sample variance as this is the most informative dimension.
  - How: formulate a constrained optimization problem and then maximize a Lagrangian. Use linear algebra to conclude that this dimension actually corresponds to the eigenvector of the largest eigenvalue.

Today's Lecture

Recap:

- Unsupervised Learning: K-means and PCA

New topic:

- Mixture modeling
- Expectation maximization; soft K-means

Mixture modeling

- Mixture modeling is useful for modeling distributions that have multiple peaks.
- Example: height distribution in a population

- Mathematically:

Gaussian Mixture Modeling (GMM)

- Can be viewed as a universal approximator for any distribution provided the number of components is sufficiently large.
- Mathematically:

- Interpretation: pick a component *k* with probability π_k. Then generate a sample according to the distribution $\mathcal{N}(\mu_k, \Sigma_k)$.

(Gaussian) Mixture Modeling

- This mixture modeling also allows for “soft” a.k.a. partial cluster membership.

- Instead of using “hard” indicators with cluster membership being strictly 0 or 1, i.e., each data point belongs to one and only one cluster, allow for fractional membership.

- Picture:

Recall: Multivariate jointly Gaussian distribution

- We have already studied this distribution.

- Where ?

- When we studied Gaussian Discriminant Analysis, both LDA and QDA.
  - This was an example of generative modeling; classification; supervised learning.

Recall: Multivariate jointly Gaussian Distribution

- Interpretation of the parameters: mean vector and covariance matrix.

Today's Lecture

Recap:

- Unsupervised Learning: K-means and PCA

New topic:

- Mixture modeling
- Expectation maximization; soft K-means

Set-up

- Say we have N data points, which we seek to organize into K clusters.
- Data likelihood:

- Goal is to figure out what the parameters are:

Set-up

- How ?

- When we encountered problems in the past that dealt with likelihood maximization w.r.t. parameters, the strategy was to write log likelihood, and set the derivatives to zero. But can’t to it here.

Expectation-maximization (EM)

- EM is a general, **iterative** method for finding a maximum (log) likelihood solution, in the presence of hidden (latent) variables.

- EM has two steps:

1. Estimates of the posteriors, called responsibilities:
   - Called responsibilities because they quantify how much is each cluster responsible in explaining a given data point.

2. MLE estimates of the parameter set:
   - Derivatives, as before.

EM – discussion

- If we knew true posteriors, we could plug them into the MLE estimates; if we knew MLE estimates, we could plug them into the posteriors.
- But we know neither!
- Solution: iterate.

- What does this approach remind you of ?

Now, for the mathematical details

- Write log likelihood:

- Define responsibilities:

Maximizing parameters

- If responsibilities were known, we can set the derivatives of the component parameters to zero.

- For the mean we get:

Maximizing parameters, ctd.

- For the mean, we get:

Maximizing parameters, ctd.

- For the mean we get:

Maximizing parameters:

- For the covariance we get:

- For the priors we get:

EM – summary

Step 1: Initialize with some values of the parameters

Step 2: Iterate between

- Computing responsibilities

- Computing parameters

- Evaluate log-likelihood – check for convergence, and terminate when appropriate

EM -- discussion

- Why is it called Expectation Maximization ?
- In the first part of every iteration, we are computing the **expectation** of Bernoulli RVs:

- In the second part of every iteration, we are **maximizing** the parameters, given the RVs from the previous part.

EM -- discussion

- Note that each half step improves on the incomplete log likelihood, so we are getting better with each iteration.

- Again, this property is reminiscent of K-means algorithm.

Connection with K-means

- Let’s consider a special case of a mixture model specified as follows:

- Recall linear algebra:

- Then, the individual distributions are:

Connection with K-means

- Then, the responsibilities are:

- Interpretation: smallest differences in the exponent dominate.

Illustrative example

Connection with K-means

- Note that the preceding derivation is essentially a principled way of performing “soft” K-means.

- Can also recover “hard” K-means as a special case: