Week4 Report

Zhengyi Wang
wangzhen17@mails.tsinghua.edu.cn

Abstract

This week, we talked about the fundemental of machine learning. Our talk is basically on *clustering* and *dimension reduction*.

1 Clustering

Clustering is one kind of the usage of unsupervised learning. Its goal is divide N items into K groups (similar items to the same group). It has a lot of different methods, including K-means and EM algorithm.

1.1 K-means

K-means algorithms is a useful method for clustering.

1.1.1 Algorithm of K-means

We define $r_{nk} \in \{0, 1\}$, and $r_{nk} = 1$ if and only if the nth item belongs to cluster k.
Our goal is to minimize the function $\sum_n \sum_k r_{nk}||x_n - \mu_k||^2$.
There are two variants, r and μ. We repeat the two step–firstly we keep μ fixed and optimize r, secondly we keep r fixed and optimize μ. Since during the optimizing process the function keeps going down, convergence is guaranteed.

1.1.2 Generalization of K-means

We can use K-means in more situations by generalization. We can define inner-product to describe the similarity between items, instead of Euclid's distance. Also, we can choose μ only from x in the cluster, for the case that the average is hard to compute.

1.2 EM algorithm

EM algoritm, short for expectation–maximization, is an outstanding algorithm for parameter estimation.

1.2.1 The algorithm of EM in details

Suppose that the data we observed are generated by a few of models (which is often the case in practice). We define a latent variable z indicating which generating model an item belongs to.
EM algorithm comprise two main steps.

- Keep other parameters fixed and find z which is most likely to render the results using MLE. (M step)
- Compute other parameters under the new z. (E step)

Reading Group on Machine Learning (May 19th, 2019).

1.2.2 Improvements of EM

In the M-step, we no longer use the MLE to compute z. Instead, we represent z by its distribution. The result will be more accurate, but the computing cost goes extremely high.

2 Dimension Reduction

Dimension reduction is useful in many cases such as data compression. The core idea is to keep the main features and leave out the others.

2.1 PCA

PCA, short for *principle component analysis*, is to project a high dimension vector onto a lower dimension. To make the information loss in this process as little as possible, we want to minimize the distance between the data and the projection. For the most striking features to be easily recognized in low dimension, we have to maximize the variance of the projection.

2.1.1 Basic Idea of PCA

We first find orthonormal bases μ for the feature space, then project each item to each of the base vector. We define sample covariance as

$$S = \frac{1}{N}\sum_{n}(x_n - \bar{x})(x_n - \bar{x})^T.$$

After some math, we find that we can find othonormal bases in the following ways–we choose the M(=the dimension after reduction) bigger eigenvalues of S and the M eigenvectors corresponding to them. The eigenvectors are our othonormal bases.

2.2 PPCA

PPCA, short for *probability PCA*, is different from PCA. We here introduce explicit latent variable z, to represent probability lying in the generating process of each item. Also useful in the case that some data are missing.

2.3 Maximum likelihood PCA

"MLPCA is a decomposition method similar to conventional PCA, but it takes into account measurement uncertainty in the decomposition process, placing less emphasis on measurements with large variance."
–cited from *Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations* Peter D. Wentzell*, Mitchell T. Lohnes
It can accommodate correlated measurement errors, but two drawbacks have limited its practical utility in these cases: (1) inability to handle rank deficient error covariance matrices, and (2) demanding memory and computational requirements.