Unsupervised Learning

Contents

You are given a some observations $\mathbf{x}_{1:n} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, e.g., a bunch of pictures, and you want to find some structure in the data. This is unsupervised learning. The key difference from supervised learning is that you do nopt have any labels/targets/outputs. The unsupervised learning problem may sound very open ended (and it is), but some specific examples we are going to study:

- Clustering: Split the observations into, say K, distinct clusters.
- Dimensionality reduction: Reduce the dimensionality of the data.
- Density estimation: Learn the probability density that gave rise to the data, i.e., learn how to generate new samples with similar features as the observations.
- and more.

Each one of these unsupervised learning problem requires several lectures to develop fully. Since we do not have this kind of time, we are going to study the most basic algorithms for the three most popular unsupervised learning problems (clustering, density estimation, and dimensionality reduction) in this and the next lecture.

Clustering using the K-means algorithm

K-means is the simplest algorithm for spliting the dataset $\mathbf{x}_{1:n}$ in K clusters. The mathematical details of the algorithm are described in the video lectures that follow. If you want study the algorithm independently, I suggest reading Chapter 20.1, D. MacKay (2003).

Density estimation using mixtures of Gaussians

As we will argue in the video, the clustering problem is related to the density estimation problem. In the density estimation problem we are trying to learn a model $p(\mathbf{x})$ that allows us to generate examples similar to our observations. We are going to work with the Gaussian mixture model which has this form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ are parameters to be estimated from the data. The Gaussian mixture model basically assumes that there are K possible sources that may generate the data and that each one of them is a multivariate Gaussian with parameters to be estimated. Solving the density estimation problem using mixtures of Gaussian essentially solves the clustering problem as well because you can think of the K different Gaussians as defining the clusters. The complete details will be developed in the video lectures. If you want to study Gaussian mixtures independently, I suggest reading Chapter 9, Bishop (2006) (You should have free online access to the book through Purdue).

Avoiding overfitting using the Bayesian information criterion

If you pick a model with too many parameters you may overfit. If you pick a model with too few parameters, you may underfit. In the Gaussian mixtures example above, you get to choose K. You can choose K to be one, and then you are essentially saying that there is a single cluster that can be descibed with a multivariate Gaussian. This may be inadequate and you are going to underfit the data. On the other extreme, you may pick K to be n, i.e., the number of observations that you have. In this case, you would be essentially fitting a different Gaussian on each observed point. This is of course ridiculous. You are definitely going to overfit.

So, how do you pick the right number of mixture components K? The formal answer is that you should be Bayesian all the way. Put a prior probability on K, the characterize the posterior over K. This is doable, it is known as Bayesian model selection (see Chapter 1.3, Bishop (2006), but we are not yet equipped to carry it out. However, there is an approximation to Bayesian model selection that is very straightforward to carrty out: the Bayesian information criterion or BIC. The details are presented in the video lecture, but if you want to study this independently the wikipedia article on BIC is well written and you may also want to take a look at Chapter 4.4.1, Bishop (2006).

By Ilias Bilionis (ibilion[at]purdue.edu)

© Copyright 2021.