VIP Cheatsheet: Unsupervised Learning

Afshine Amidi and Shervine Amidi

August 5, 2018

Introduction to Unsupervised Learning

❒ **Motivation** – The goal of unsupervised learning is to find hidden patterns in unlabeled data $\{x^{(1)},...,x^{(m)}\}$.

❒ **Jensen's inequality** – Let f be a convex function and X a random variable. We have the following inequality:

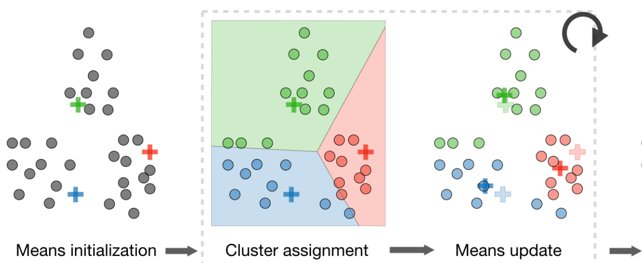
$$\boxed{E[f(X)]\geqslant f(E[X])}$$

k-means clustering

We note $c^{(i)}$ the cluster of data point i and μ_j the center of cluster j.

❒ **Algorithm** – After randomly initializing the cluster centroids $\mu_1,\mu_2,...,\mu_k\in\mathbb{R}^n$, the k-means algorithm repeats the following step until convergence:

$$\boxed{c^{(i)}=\underset{j}{\textrm{arg min}}||x^{(i)}-\mu_j||^2}\quad\textrm{and}\quad\boxed{\mu_j=\frac{\displaystyle\sum_{i=1}^m1_{\{c^{(i)}=j\}}x^{(i)}}{\displaystyle\sum_{i=1}^m1_{\{c^{(i)}=j\}}}}$$



❒ **Distortion function** – In order to see if the algorithm converges, we look at the distortion function defined as follows:

$$\boxed{J(c,\mu)=\sum_{i=1}^m||x^{(i)}-\mu_{c^{(i)}}||^2}$$

Expectation-Maximization

❒ **Latent variables** – Latent variables are hidden/unobserved variables that make estimation problems difficult, and are often denoted z. Here are the most common settings where there are latent variables:

Setting	**Latent variable z**	$x\|z$	**Comments**
Mixture of k Gaussians	Multinomial(ϕ)	$\mathcal{N}(\mu_j,\Sigma_j)$	$\mu_j\in\mathbb{R}^n,\phi\in\mathbb{R}^k$
Factor analysis	$\mathcal{N}(0,I)$	$\mathcal{N}(\mu+\Lambda z,\psi)$	$\mu_j\in\mathbb{R}^n$

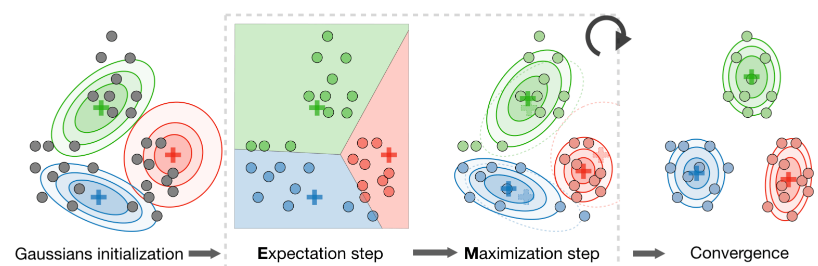
❒ **Algorithm** – The Expectation-Maximization (EM) algorithm gives an efficient method at estimating the parameter θ through maximum likelihood estimation by repeatedly constructing a lower-bound on the likelihood (E-step) and optimizing that lower bound (M-step) as follows:

- E-step: Evaluate the posterior probability $Q_i(z^{(i)})$ that each data point $x^{(i)}$ came from a particular cluster $z^{(i)}$ as follows:

$$\boxed{Q_i(z^{(i)})=P(z^{(i)}|x^{(i)};\theta)}$$

- M-step: Use the posterior probabilities $Q_i(z^{(i)})$ as cluster specific weights on data points $x^{(i)}$ to separately re-estimate each cluster model as follows:

$$\boxed{\theta_i=\underset{\theta}{\textrm{argmax }}\sum_i\int_{z^{(i)}}Q_i(z^{(i)})\log\left(\frac{P(x^{(i)},z^{(i)};\theta)}{Q_i(z^{(i)})}\right)dz^{(i)}}$$



Principal component analysis

It is a dimension reduction technique that finds the variance maximizing directions onto which to project the data.

❒ **Eigenvalue, eigenvector** – Given a matrix $A\in\mathbb{R}^{n\times n}$, λ is said to be an eigenvalue of A if there exists a vector $z\in\mathbb{R}^n\backslash\{0\}$, called eigenvector, such that we have:

$$\boxed{Az=\lambda z}$$

❒ **Spectral theorem** – Let $A\in\mathbb{R}^{n\times n}$. If A is symmetric, then A is diagonalizable by a real orthogonal matrix $U\in\mathbb{R}^{n\times n}$. By noting $\Lambda=\textrm{diag}(\lambda_1,...,\lambda_n)$, we have:

$$\boxed{\exists\Lambda \textrm{ diagonal},\quad A=U\Lambda U^T}$$

Remark: the eigenvector associated with the largest eigenvalue is called principal eigenvector of matrix A.

❒ **Algorithm** – The Principal Component Analysis (PCA) procedure is a dimension reduction technique that projects the data on k dimensions by maximizing the variance of the data as follows:

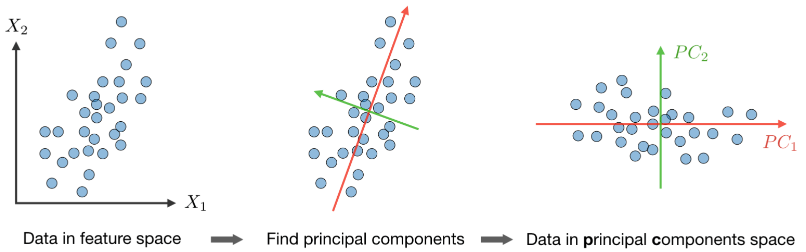
- Step 1: Normalize the data to have a mean of 0 and standard deviation of 1.

$$\boxed{x_j^{(i)}\leftarrow\frac{x_j^{(i)}-\mu_j}{\sigma_j}}\quad\textrm{where}\quad\boxed{\mu_j=\frac{1}{m}\sum_{i=1}^m x_j^{(i)}}\quad\textrm{and}\quad\boxed{\sigma_j^2=\frac{1}{m}\sum_{i=1}^m(x_j^{(i)}-\mu_j)^2}$$

- Step 2: Compute $\displaystyle\Sigma=\frac{1}{m}\sum_{i=1}^m x^{(i)}{x^{(i)}}^T\in\mathbb{R}^{n\times n}$, which is symmetric with real eigenvalues.

- Step 3: Compute $u_1,...,u_k\in\mathbb{R}^n$ the k orthogonal principal eigenvectors of Σ, i.e. the orthogonal eigenvectors of the k largest eigenvalues.

- Step 4: Project the data on $\textrm{span}_\mathbb{R}(u_1,...,u_k)$. This procedure maximizes the variance among all k-dimensional spaces.



Independent component analysis

It is a technique meant to find the underlying generating sources.

❒ **Assumptions** – We assume that our data x has been generated by the n-dimensional source vector $s=(s_1,...,s_n)$, where s_i are independent random variables, via a mixing and non-singular matrix A as follows:

$$\boxed{x=As}$$

The goal is to find the unmixing matrix $W=A^{-1}$ by an update rule.