1 Representation Learning

Representation Learning is a a group of methods that transform the data into relevant features.

1.1 Basics of Representation Learning

- Performance of Machine Learning methods can be highly dependent on how the data is represented
- Representation Learning seeks to learn a transformation of the original data so that classic machine learning methods can be effectively utilized
- Data types such as images and videos can often be redundant and highly variable, so it is of interest to transform the data in a manner to see useful features
- These algorithms can be supervised or unsupervised
- In unsupervised methods, the goal is to estimate relevant features of p(X), the joint distribution of X
- K-means Clustering can be used to find k centroids from unlabelled data which can then be used to produce k features.
- In supervised methods, we form feature maps that take into account the joint distribution p(X,Y)

1.2 Representation Paradigms

- Probabilistic Methods
- Auto-encoders
- Manifold Learning

1.3 Principal Component Analysis

- Principal Component Analysis (PCA) is an unsupervised method that can result in dimension reduction.
- PCA solves a variety of optimization problems.
- If we want to find the first q principal components, the relevant optimization program is:

$$\min_{\mu,(\lambda_i),V_q} \sum_{i=1}^{n} ||X_i - \mu - V_q \lambda_i||^2$$

- Principal components can be viewed as coming from all three representation learning paradigms.
- Several examples using images are presented in the class notes.

1.4 Sparce Coding

- Sparce coding or neural coding is based on the idea that a network of neurons in the brain codes visual information in a particular manner
- We possess a basis of neurons that permits certain types of images to be expressed sparsely
- In this particular context, the sparceness comes from only a few non-zero coefficients
- We begin by assuming that we have a dictionary $\Phi \in R^{pxK}$ with K ¿ p

1.5 Basis Pursuit

Let $\Phi = [\phi_1, \phi_2, ..., \phi_k]$. Rather than having a set of covariates in the LASSO, we can think of the covariates being a basis. Our goal is then to minimize the following.

$$\min_{\alpha} ||Y - \Phi\alpha||_2^2 + \lambda||\alpha||_1$$

One example of such a basis would be a combination of Fourier basis and a Wavelet basis. In particular, the span of this set is NOT linearly independent.

1.6 More on Sparce Coding

The tuning parameter λ can be set

$$\lambda_* = \sigma\sqrt{2\log(K)}$$

The problem can be now be rewritten:

$$\min_{\Phi, \alpha \in \mathbb{R}^{k \times n}} \sum_{i=1}^{n} (||X_i - \Phi\alpha_i||_2^2 + \lambda||\alpha_i||)$$
$$subjectto||\Phi|| \leq c$$

- A stochastic gradient method approach can work well.
- Finding α is not too difficult using a lasso-type procedure
- Finding Φ is more complex. A classical, fast approach is a projected first-order stochastic gradient descent method.
- Details on this method are available online.
- Examples of sparce coding were examined in the class notes.

Deep Learning will be examined in the next set of notes.