Machine Learning 1 - Homework assignment 4

Available: Monday, November 26th, 2018
Deadline: Thursday 23.59, December 6th, 2018

General instructions
Unless stated otherwise, write down a derivation of your solutions. Solutions presented without a derivation that shows how the solution was obtained will not be awarded with points.

1 Mixture of Experts

In class you discussed and were introduced to mixture models as a way to perform unsupervised learning tasks, e.g. clustering. Mixture models are not limited to only unsupervised learning and can be similarly used for supervised learning. In this homework we will discuss and explore Mixtures of Experts (MoEs), a model that softly partitions the input space and learns a supervised model for each area.

Consider that you have K experts available in order to model a specific dataset of N datapoints $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_n$ corresponds to a vector input in $\mathbb{R}^D$ and y_n corresponds to the particular label available for $\mathbf{x}_n$. Let z_n correspond to a categorical random variable for datapoint n that denotes which of the K experts is active. Furthermore, let $\mathbf{\Theta}$ be a matrix in $\mathbb{R}^{D\times K}$ that contains the D-dimensional column vector of parameters for each expert. We will assume that each y_i is a continuous random variable at the $[0, \infty)$ interval distributed according to an exponential distribution with a rate $\lambda > 0$. Given the aforementioned assumptions, each expert $k \in K$ has the following linear predictive model:

$$\begin{aligned} p(y_n|\mathbf{x}_n, z_n, \mathbf{\Theta}) &= p(y_n|\mathbf{x}_n, \boldsymbol{\theta}_k = \mathbf{\Theta}\mathbf{z}_n), \\ &= \text{Exponential}(y_n|\lambda = \exp(\boldsymbol{\theta}_k^T\mathbf{x}_n)), \end{aligned}$$

where $\mathbf{z}_n$ corresponds to a 1-of-K vector representation of the categorical variable z_n and

$$\text{Exponential}(y|\lambda) = \lambda\exp(-\lambda y) \text{ for } y \geq 0.$$

The flexibility of MoEs stem from the fact that there is a “routing” mechanism which determines which of the K experts is appropriate for a specific datapoint $\mathbf{x}_n$. As in this case we have a discrete set of K experts, a simple linear routing mechanism is the following:

$$p(z_n = k|\mathbf{x}_n, \mathbf{\Phi}) = \pi_{nk} = \frac{\exp(\boldsymbol{\phi}_k^T\mathbf{x}_n)}{\sum_j \exp(\boldsymbol{\phi}_j^T\mathbf{x}_n)},$$

where $\mathbf{\Phi}$ is a matrix in $\mathbb{R}^{D\times K}$ that contains all of the parameters of the routing function, i.e. $\mathbf{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K]$. As a-priori we have no information about which of the experts is responsible for generating a particular prediction we have to marginalize over all possible experts in order to compute the likelihood of an observed point.

With this information answer the following questions:

1. Write down the likelihood of the entire dataset, $p(\mathbf{y}|\mathbf{X}, \mathbf{\Theta}, \mathbf{\Phi})$, and take its log under the i.i.d. assumption. (**1 pt.**)
2. Write down the posterior probability r_{ni} of expert i producing the label y for datapoint n. We will also refer to this as the responsibility of expert i for datapoint n. (**1 pt.**)
3. Take the derivative of the log-likelihood w.r.t. the parameters of each expert $\boldsymbol{\theta}_i$ and the parameters of the routing mechanism for each expert $\boldsymbol{\phi}_i$. Do not substitute expressions for the probabilities but rather provide your answer in terms of $p(y_n|\mathbf{x}_n, z_n, \boldsymbol{\theta}_i)$, $p(z_n = k|\mathbf{x}_n, \mathbf{\Phi})$. Make sure to express the derivatives in terms of the responsibilities of each expert r_{ni}. (Hint: $\frac{\partial f(x)}{\partial x} = f(x)\frac{\partial \log f(x)}{\partial x}$), as that term will be present in the derivatives for both $\boldsymbol{\theta}_i, \boldsymbol{\phi}_i$. (**4 pt.**)
4. Replace the expressions for each of the respective probability distributions and compute the final derivatives for $\boldsymbol{\theta}_i, \boldsymbol{\phi}_i$. (**4 pt.**)
5. Write down an iterative algorithm that maximizes the log-probability of the data by jointly optimizing the $\mathbf{\Theta}$ and $\mathbf{\Phi}$ parameters. Make use of appropriate convergence criteria. (**1 pt.**)

Assume that an oracle is available that provides you an extra set of M points that also has the information about which expert t should be employed.

1. Write down the likelihood of this extended dataset and take its log. (**1 pt.**)
2. Take the derivatives of this new log-likelihood w.r.t the parameters $\boldsymbol{\theta}_i, \boldsymbol{\phi}_i$. (**2 pt.**)
3. What is the difference between the derivatives computed here and the derivatives computed previously? Is the overall model a linear or a non-linear one? (**2pt.**)

2 Principal Component Analysis

Suppose we have a data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of D-dimensional vectors, which have a zero mean for each dimension. Assume we perform a complete eigenvalue decomposition of the empirical covariance matrix $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$.

1. Initially, you are interested in only a single projection of your data such that the variance of this projection is maximized. Let $\mathbf{u}_i$ be the direction vector of a particular projection. Assume that $\mathbf{u}_i^T\mathbf{u}_i = 1$.
   (a) What is the projection z_{ni} of a given point $\mathbf{x}_n$ under the particular vector $\mathbf{u}_i$? (**1 pt.**)
   (b) What is the empirical mean of the projection z_i across all points $\mathbf{x}_n$? (**1 pt.**)

(c) What is the empirical variance of the projection z_i? Provide your answer in terms of the empirical covariance matrix $\mathbf{S}$. (**1 pt.**)

(d) Replace $\mathbf{S}$ with its eigenvalue decomposition and simplify the aforementioned expression. What is the variance now? (**2 pt.**)

(e) Suppose that you are interesting in reducing the dimensionality from D to K, such that 99% of the variance is maintained. How can you select an appropriate K? (**2 pt.**)

2. Consider the projections of your data along the K principal components. Prove that the dimensions of the projections are de-correlated (Hint: check the value of the empirical covariance w.r.t. dimension i and j). (**2 pt.**)

3. Imagine that you want to de-correlate all of the dimensions but still want to enforce a mean of $\mathbf{m}$ and variance of τ across the D dimensions. How can you post-process your projections such that they satisfy these properties? Show that these hold by computing the empirical mean and variance across the dataset. (**3 pt.**)