Machine Learning 1 Homework Week 6

Monday, October 5, 2015
Deadline: Friday, October 17, 2015, 23:59

1 Decision Trees

We are given the following dataset: $\{\mathbf{x}_i, y_i\}$, $i = 1..N$, where x_{ai} is the a'th attribute ($a = 1..A$) of the i'th data-case and $y_i = \{-1, +1\}$. Each x_a can take one of 2 discrete values $\{A, B\}$.

(a) Assume that you have partially trained a decision tree and are considering to add another attribute to the tree at some branch. At this point you may assume you have n (negative) data cases in class -1 and p (positive) data cases in class $+1$. Moreover, after you apply the attribute, you will have n_A negative data cases with attribute value A, n_B negative data cases with attribute value B, p_A positive data cases with attribute value A and p_B positive data cases with attribute value B. Provide the expression for the *information gain* for this attribute.

(b) At some point you find that in some branch you have used all your attributes, but that there are still some positive and negative data-cases that did not get resolved (or split). Say there are p_L positive and n_L negative data-cases left in this leaf L. Consider now a test case that ends up at this leaf node of this branch in the decision tree. What procedure would you follow to classify this test case as either positive or negative?

(c) Describe how the random forest algorithm works (based on decision trees). Again, use pseudo-code style to explain your answer.

2 PCA and kernel-PCA

Suppose we have a dataset of N vectors $\{\mathbf{x}_n\}$ of dimension D. We can write the entire dataset as a D by N matrix $\mathbf{X}$ (column n is x_n). We may wish to perform PCA on this data in the original data space, or in *kernel*-space using kernel-PCA. In the latter case, the data are projected into *feature* space $\boldsymbol{\phi}$, such that $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ is M-dimensional feature space representation of x_n. Consider the procedure for PCA (which can be generalized to kernel-PCA):

Step 1 Center $\mathbf{X}$, producing a center data matrix $\hat{\mathbf{X}}$.

Step 2 Compute sample covariance $\mathbf{S}$ of the centered dataset.

Step 3 Solve the eigen-value problem $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\mathbf{U}$ is a column matrix of eigen-vectors and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigen-values λ_k, ie $\boldsymbol{\Lambda}_{kl} = \lambda_k \delta_{kl}$, where $\delta_{kl} = 1$ iff $k = l$.

Step 4 Pick eigen-vectors with largest eigen-values $\{\mathbf{u}_1, \ldots \mathbf{u}_K\}$.

Step 5 Project data onto K-dimensional manifold.

Answer the following questions:

(a) Provide an expression for $\hat{\mathbf{x}}_n$.

(b) Prove that the average of $\hat{\mathbf{x}}_n$ (over N data vectors) is the 0 vector.

(c) Provide an expression for $\mathbf{S}$ in terms of $\hat{\mathbf{X}}$.

(d) What is the dimensionality of $\mathbf{S}$?

(e) What is the expression for the linear projection $\mathbf{L}$ that maps data vectors $\hat{\mathbf{x}}_n$ onto a K-dimensional sub-space, $y_n = \mathbf{L}\hat{\mathbf{x}}_n$, such that it has zero mean and identity covariance. Prove that the average over N of y_n is 0. Prove that the covariance of y_n is the identity. What is this operation called?

(f) For kernel-PCA, the centering step cannot in general be performed in the feature space.

  (i) Write the equation for centered feature vector $\hat{\boldsymbol{\phi}}_n$.

  (ii) The eigen-vector problem solved in kernel-PCA uses a kernel matrix $\hat{\mathbf{K}}$ that is centered in kernel-space. Using your result above for the center feature vector, expand the gram matrix entry $\hat{\mathbf{K}}_{nm} = \hat{\boldsymbol{\phi}}_n^T \hat{\boldsymbol{\phi}}_m$ in terms of $\mathbf{k}$ only (ie in terms of the non-centered kernel functions).

(g) In terms of memory requirements, what is the disadvantage of kernel-PCA versus PCA?

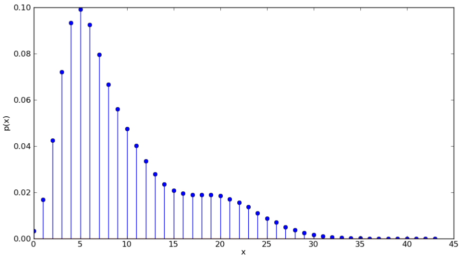

Figure 1: The empirical distribution of data drawn from a mixture of $K = 3$ Poisson distributions.

3 Mixture Models

Consider the data distribution shown in Figure 1. Each vertical line represents the empirical probability distribution of a dataset of discrete data values x (the frequency of a particular value x in the dataset). We are told that the generating process is a mixture of Poisson distributions, but we do not know the parameters of the mixture model. In this question you are asked to derive the update equations for the general Poisson mixture model.

The Poisson distribution is:

$$P(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)$$

where $x = 0, 1, 2, ...$ (non-negative integers), $\lambda > 0$ is the 'rate' of the data; the expected value of x is λ. A mixture representation assumes the following:

$$P(x_n) = \sum_{k=1}^{K} \pi_k P(x_n|\lambda_k)$$

where $P(x_n|\lambda_k)$ is a Poisson distribution with rate λ_k and x_n is a single data observation. To answer the following questions assume we are given a dataset $\{x_1, x_2, \ldots, x_N\}$. Make sure that the constraint $\sum_k \pi_k = 1$ is satisfied (i.e. think of the log-likelihood or log-joint as f (an objective to maximize) and $\sum_k \pi_k - 1 = 0$ as $g = 0$ (a constraint that must hold)).

(a) Write down the likelihood (as usual) for the data set in terms of $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.

(b) Write down the log-likelihood (as usual) for the data set in terms of $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.

(c) Find the expression for the responsibilities r_{nk}.

(d) Find the expression for λ_k that maximizes the log-likelihood.

(e) Find the expression for π_k that maximizes the log-likelihood.

(f) Now assume priors for π_k and λ_k. $p(\lambda_k|a, b) = \mathcal{G}(\lambda_k|a, b)$ (a Gamma prior) and $p(\pi_1, \ldots, \pi_k) = \mathcal{D}(\pi_1, \ldots, \pi_k|\alpha/K, \ldots, \alpha/K)$ (a Dirchlet distribution). These distributions are defined in the appendix of Bishop. Write down the log-joint distribution
$\log p(\mathbf{x}_1, \ldots, \mathbf{x}_N, \{\pi_k\}, \{\lambda_k\}|a, b, \alpha, K)$.

(g) Find the expression for λ_k that maximizes the log-joint.

(h) Find the expression for π_k that maximizes the log-joint.

(i) Write down an iterative algorithm using the above update equations (similar to the ones derived in class for the Mixture of Gaussians); include initialization and convergence check steps.