Course 02457 Non-Linear Signal Processing, Exercise 9

This exercise is based on C. M. Bishop: *Pattern Recognition and Machine Learning*, section 2.5.

Print and comment on the figures produced by the software as outlined below at the **Checkpoints**.

Probability density function estimation using a kernel smoother

A training set of N data points $D = \{\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_N\}$ is extrapolated 'smoothened' to test points $\mathbf{x}$

$$p(\mathbf{x}|D,h) \quad = \quad \frac{1}{N}\sum_{n=1}^{N} k(\mathbf{x}|\mathbf{x}_n,h)$$

with a 'kernel' $k(\mathbf{x}|\mathbf{x}_n,h)$ given eg. by

$$k(\mathbf{x}|\mathbf{x}_n,h) \quad = \quad \left(\frac{1}{2\pi h^2}\right)^{d/2} \exp\left(-\frac{1}{2h^2}(\mathbf{x}-\mathbf{x}_n)^2\right)$$

where the dimension of $\mathbf{x}$ is d. The parameter h acts as a smoothing control. If h is small we roughly get a set of local 'delta functions' centered on the training data set, if h, on the other hand, is very large we get a near-uniform distribution.

Checkpoint 9.1

We will use a validation set - a test set for tuning of parameters - of M samples to find h. Explain why the function

$$E(h) = \frac{1}{M}\sum_{m=1}^{M} -\log p(\mathbf{x}_m|D,h) \quad = \quad \frac{1}{M}\sum_{m=1}^{M} -\log \frac{1}{N}\sum_{n=1}^{N} k(\mathbf{x}_m|\mathbf{x}_n,h)$$

is a 'test error'. Use the matlab script `main9a.m` to generate data from a normal distribution in $d = 2$. What is the optimal h for this data set. Explain the structure of the densities obtained by the 'optimal' h, and h's that are too small and too big.

How does the optimal h depend on the training sample size N?

Signal detection using nearest neighbor methods

We will use nearest neighbor methods for non-parametric classification. Assume that a training set with N class labeled samples is given.

In the K-nearest-neighbor (KNN) classifier we classify test points $\mathbf{x}$ by voting among the K nearest neighbors in the training set. We implement this by brute force, i.e., simply by computing the distance from the test point to all training points and sorting the distances.

A so-called 'leave one out' estimate of the classification test error can be obtained by computing the distances from every training point to its K neighbors (not including itself!) and in turn estimate the classification error of the voting result among the neighbors relative to the given training point's label.

Pima indian data set

This is a data set where the task is to classify a population of women according to the risk of diabetes (binary classification). There are 7 input variables, 200 training examples and 332 test examples. 68 (34%) in the training set and 109 (32.82%) in the test set have been diagnosed with diabetes. In Brian Ripley's textbook *Pattern Recognition and Neural Networks* he states that his best method obtains about 20% misclassifications on the test data set. The input variables are:

1. Number of pregnancies
2. Plasma glucose concentration
3. Diastolic blood presure
4. Triceps skin fold thickness
5. Body mass index (weight/height^2)
6. Diabetes pedigree function
7. Age

The target output is 1 for examples diagnosed as diabetes, and 2 for healthy subjects.

Checkpoint 9.2

Explain how the 'leave one out' error can be used for identifying the optimal number of neighbors for voting. Use the matlab script `main9b.m` to classify the diabetes diagnosis data set. What is the optimal K? How well is the KNN performance compared to neural networks and other methods considered earlier in the course.

Consider classification from a subset of the seven input variable measures. Estimate the performance for a few subsets, can you find a subset with performance equal or better than that of the full feature set?

Local linear regression among nearest neighbors

We can design a non-parametric function approximation scheme by performing linear regression among the K-nearest neighbors. We apply the method to prediction of the sunspot test data set.

We use the linear model from exercise 3 and 4 to perform the estimation in the test set.

Checkpoint 9.3

Inspect the matlab script `main9c.m`. Explain the role of the parameter 'alpha'. Why is it necessary to regularize the linear model? What is the meaning of the parameter d and what is optimal value of d. Make a drawing that explains the algorithm conceptually, e.g., in a case with two-dimensional input.

Compare the quality of the algorithm's predictions with the neural network based predictions we found in exercise 5. What would happen if we used $K = N_{train}$?.

Challenge (not part of the curriculum)

Inspect the local linear models in Checkpoint 9.3 for test points with respect to mean, variance and 'outliers' (e.g., order test points according to how far away their local linear models are from the mean). Where are the test points with 'outlier models', located in the sun-spot data?. Are the outliers consistent with respect to models trained with different 'K' and 'alpha'?

DTU, November 2007, 2013, Lars Kai Hansen