COURSE 02457

Non-Linear Signal Processing: Exercise 8

This exercise is based on in C. M. Bishop: *Pattern Recognition and Machine Learning* section 6.3.

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

Regression based on density estimation

We observe a stochastic multi-channel signal with d-dimensional inputs $\mathbf{x}$ and scalar outputs t and our aim is model the density $p(\mathbf{x}, t) \sim p(\mathbf{x}, t|\mathbf{w})$ where the family $p(\mathbf{x}, t|\mathbf{w})$ is a given parametric density. The training set is consists of N input-output pairs.

The mixture of Gaussians model family is defined

$$p(\mathbf{x}, t|\mathbf{w}) = \sum_{j=1}^{M} p(\mathbf{x}, t|\mathbf{w}_j) P(j) \ ,$$

where each of the M component densities is a normal distribution with parameters $\mathbf{w}_j = \left\{\boldsymbol{\mu}_j, \sigma_j^2\right\}$. $P(j)$ is the mixing proportion for component j, that is the prior probability that the data comes from this component. Here we will invoke a family of "isotropic" Gaussians, i.e., Gaussians with covariance matrices that are scaled unit matrices,

$$p(\mathbf{x}, t|\boldsymbol{\mu}_j, \sigma_j^2) \ = \ \frac{1}{(2\pi\sigma_j^2)^{(d+1)/2}} \exp\left(-\frac{||\mathbf{x} - \boldsymbol{\mu}_j||^2 + (t - \nu_j)^2}{2\sigma_j^2}\right) .$$

$\boldsymbol{\mu}_j$ is mean vector in input space and ν_j the mean in the output space. This model makes the important simplifying assumption that within each component, the variance is identical in all directions including the output direction.

We will estimate the parameters using maximum likelihood as implemented by the Expectation-Maximization algorithm as in exercise 7 (see also Bishop chapter 9). The cost function is therefore

$$E = -\sum_{n=1}^{N} \log p(\mathbf{x}_n, t_n|\mathbf{w}) \ .$$

Based on the joint input-output distribution we can compute the conditional distribution

$$\begin{aligned} p(t|\mathbf{x}) \ &= \ \frac{p(\mathbf{x}, t)}{p(\mathbf{x})} \\ p(\mathbf{x}) \ &= \ \int p(\mathbf{x}, t)\, dt = \sum_{j'=1}^{M} \frac{P(j')}{(2\pi\sigma_{j'}^2)^{\frac{d}{2}}} \exp\left(-\frac{||\mathbf{x} - \boldsymbol{\mu}_{j'}||^2}{2\sigma_{j'}^2}\right) \ , \end{aligned}$$

where the explicit expression for $p(\mathbf{x})$ obtained by marginalization over t is a standard Gaussian integral (see Bishop Section 2.3). The conditional output mean (Bishop section 6.3) is the optimal regression function for least squares loss function:

$$y(\mathbf{x}) \;\; = \;\; < t|\mathbf{x} > \equiv \int t\, p(t|\mathbf{x}) dt \; .$$

The explicit expression for the conditional mean is

$$y(\mathbf{x}) = \frac{\sum_{j=1}^{M} \nu_j \, \frac{P(j)}{(2\pi\sigma_j^2)^{\frac{d}{2}}} \exp\left(-\frac{||\mathbf{x}-\boldsymbol{\mu}_j||^2}{2\sigma_j^2}\right)}{\sum_{j'=1}^{M} \frac{P(j')}{(2\pi\sigma_{j'}^2)^{\frac{d}{2}}} \exp\left(-\frac{||\mathbf{x}-\boldsymbol{\mu}_{j'}||^2}{2\sigma_{j'}^2}\right)}$$

We see that this expression contains the responsibility

$$P(j|\mathbf{x}) = \frac{\frac{P(j)}{(2\pi\sigma_j^2)^{\frac{d}{2}}} \exp\left(-\frac{||\mathbf{x}-\boldsymbol{\mu}_j||^2}{2\sigma_j^2}\right)}{\sum_{j'=1}^{M} \frac{P(j')}{(2\pi\sigma_{j'}^2)^{\frac{d}{2}}} \exp\left(-\frac{||\mathbf{x}-\boldsymbol{\mu}_{j'}||^2}{2\sigma_{j'}^2}\right)} \; .$$

The conditional mean is therefore a convex combination of the components means in output space:

$$y(\mathbf{x}) = \sum_{j=1}^{M} \nu_j \, P(j|\mathbf{x}) \; .$$

Checkpoint 8.1

Use the program `main8a.m` to perform EM learning of the parameters for the sunspot data. The program creates four figures. The first figure is a scatter plot of training points $x(k)$ versus $y(k) = x(k+1)$, and also the location of the centers as iterations progress. The second plot shows the evolution of the variances σ_j. By default we clamp them to be identical by setting `common-sigs = 1`. The second figure also shows the evolution of the training and test errors of the density estimates. The third figure is like figure 1, but with the final clusters. Figure 4 shows the training and test set time series. The program also reports (in the Matlab prompt line) the training and test errors estimated by prediction. These are calculated as in earlier exercises by the normalized squared error of the predictions on training and test sets. In the program you can also change the number of mixture components K. First run the program with $K = 5, 25$ with common covariances. Do you see overfitting in the density test error (figure 2)? Do you see overfitting in the squared error? Set `common-sigs = 0` and comment on the results for $K = 5, 25$. If you experience problems with shrinking variances, try to introduce a lower bound, what is a good value? Compare the results obtained with those of multilayer perceptron nets from earlier exercises.

`Test test test`

Checkpoint 8.2

To run the program `main8a.m` with higher K values you may want to eliminate intermediate plots by setting `plot-motion=0`. Run the algorithm for higher K's to find the best K from a prediction point of view.

Signal detection based on density estimation

If we adapt density models $P(\mathbf{x}|C_k)$ for each class of a classification problem we can use Bayes' theorem to obtain the posterior probabilities (see Bishop section 1.5 and this week's lecture slide),

$$P(C_k|\mathbf{x}) \quad = \quad \frac{\sum_j p(\mathbf{x}|j)P(j|C_k)P(C_k)}{\sum_{j'} p(\mathbf{x}|j')P(j')} \; .$$

In the above expression we use a mixture of Gaussians to model the density, i.e. $p(\mathbf{x}|j)$ is a normal and $P(j|C_k)$ is probability (table) indicating which components belong to which classes.

Checkpoint 8.3

Use the program `main8b.m` to adapt densities for the two classes of the "Pima indian problem". In figure 1 we show scatter plot of two of the seven pima inputs and superimpose the motion of the centers as we adapt the densities. This program also reports two test errors: a density estimate test error (figure 2) and the classification test error (prompt line). Figure 3 shows the location of the components for the final configuration. Create a flowchart description of the program. Run the program for different K, comment on the location of the components. Which value of K can you recommend?

Challenge I (not part of the curriculum)

Use the results of Challenge I in Exercise 7 to run the Checkpoints in this exercise with full covariance matrices instead of isotropic covariances. Do you see improvements over the isotropic model results?

DTU, 2006 Lars Kai Hansen (2007, 2009 Ole Winther)