Introduction to Intelligent Systems

Written Exam 26-10-2009

Duration: **3 hours. No additional material allowed – this is a closed book exam!**

✓ **1) (1pt) LVQ**

Consider a set $D = \{\boldsymbol{\xi}^\mu, S^\mu\}_{\mu=1}^{P}$ containing N-dim. feature vectors $\boldsymbol{\xi}^\mu$ and labels $S^\mu \in \{1,2\}$. Explain the supervised Learning Vector Quantization algorithm (LVQ1) discussed in class. Restrict the discussion to the simple Euclidean distance measure and the use of two N-dimensional prototypes $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ (one for each class).

Explain the algorithm in terms of a few lines of pseudocode. Be precise and provide equations, e.g. defining the *winning prototype* or the actual update step. Explain the parameters of the algorithm. In which order you would present the data and why?

✓ **2) (1 pt) Distance measures**

Consider an LVQ classifier with two prototypes in two dimensions, i.e. inputs $\boldsymbol{\xi} \in I\!R^2$. Assume the prototypes representing class 1 and 2 are located in $\boldsymbol{w}^{(1)} = (0,1)^\top$ and $\boldsymbol{w}^{(2)} = (1,0)^\top$, respectively. Give an analytical expression, of the decision boundary (i.e. an equation relating ξ_1 and ξ_2 for points on boundary) for the following distance measures. Also sketch the respective decision boundaries graphically.

a) squared Euclidean distance, i.e.

$$d(\boldsymbol{w}^{(j)}, \boldsymbol{\xi}) = \sum \left(w_i^{(j)} - \xi_i\right)^2$$

b) modified Euclidean distance with global relevances, i.e.

$$d(\boldsymbol{w}^{(j)}, \boldsymbol{\xi}) = \sum_{i=1}^{2} \lambda_i \left(w_i^{(j)} - \xi_i\right)^2 \quad \text{for } \lambda_1 = 0.2, \lambda_2 = 0.8$$

c) modified Euclidean distance with local relevances, i.e.

$$d(\boldsymbol{w}^j, \boldsymbol{\xi}) = \sum_{i=1}^{2} \lambda_i^{(j)} \left(w_i^{(j)} - \xi_i\right)^2 \quad \text{for } \lambda_1^{(1)} = 0.2, \lambda_2^{(1)} = 0.8 \quad \text{and } \lambda_1^{(2)} = \lambda_2^{(2)} = 0.5$$

Hint: the decision boundary is given by the set of points that have equal distance from $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$.

✓ **3) (1 pt) Cross-Validation**

Consider a classification problem for which a set of P labeled examples is given. Explain in your own words the method of *k-fold cross validation*. How can it be used to estimate the generalization quality of a classifier? How can it be used for model selection and optimization of algorithm parameters, e.g. the learning rate in LVQ?

Which method do you obtain for $k = P$ in k-fold cross validation? What are the advantages and disadvantages of this extreme case?

4) (1 pt) **Bayesian classification. Normal distributions**

Let us consider a two-category classification problem, with categories A and B with equal prior probabilities $P_A = P_B$. The class-conditional probability densities $p_{x|A}$ and $p_{x|B}$ are one-dimensional normal distributions:

$$p_{x|A} \sim \mathcal{N}(\mu_A, \sigma_A^2), \qquad p_{x|B} \sim \mathcal{N}(\mu_B, \sigma_B^2).$$

✓a) Express the position(s) of the optimal Bayesian decision boundary or boundaries as a function(s) of $\mu_A, \sigma_A, \mu_B, \sigma_B$.

✓b) Under which condition there is exactly 1 decision boundary.

c) Let us consider the sets of observations $\{-3, -2, -1, 0, 1\}$ for category A and $\{2, 3, 4, 5, 6\}$ for category B.

c1) Compute maximum likelihood estimates of $\mu_A, \sigma_A, \mu_B, \sigma_B$.

c2) Compute the position(s) of the decision boundary (or boundaries).

✓ 5) (1 pt) **Hierarchical clustering**

Consider the following set of numbers $S = \{3, 4, 6, 10, 12, 13, 16\}$. The dissimilarity between two numers is defined as the absolute value of their difference. The dissimilarity between two clusters is defined by the dissimilarity of their least dissimilar elements (single linkage algorithm).

a) Build a dendrogram for this set.

b) Using the dendrogram, cluster the numbers in two clusters.

c) Using the dendrogram, cluster the numbers in three clusters.

Hint: Using the dissimilarity matrix approach will take you a lot of time. You can build the dendrogram faster if you plot the data on the 1D real number axis and decide visually how to cluster data agglomeratively.

✓6) (1 pt) **k-nearest neighbors classification**

Consider the following two sets of points in a 2D space:

$$A = \{(2,3), (3,2), (3,3), (3,4), (3,5), (4,3), (5,3)\}$$
$$B = \{4,4), (4,6), (5,6), (6,6), (6,5), (6,7), (7,6)\}$$

which are drawn from two different classes ω_A and ω_B, respectively. Using Euclidean distance and the 3-nearest neighbors algorithm, classify the following test points: $(6,4), (6,3), (2,6), (4,5)$.

Hint: Computing distances may take you a lot of time. You can solve the problem faster by plotting the data and deciding visually which the three nearest neighbors of a test point are.

-1	0	1
-1	0	1
-1	0	1

0	-1	0
0	2	0
0	-1	0

Figure 1: *Convolution masks for the Sobel x-gradient filter (left) and the second-order y-derivative filter (right).*

√ 7) (1.5 pt) **Edge Detection.**
Consider a grey-value image f.

a. (0.5pt) Prewit gradients in the image in horizontal (easterly) direction can be detected by linear filtering using the filter kernel (or mask) in Fig. 1 (left). Give the Prewit kernel to detect gradients in northerly direction.

b. (1 pt) A discrete second derivative filter in the x-direction $\frac{\partial^2}{\partial y^2}$ is defined by convolution with the kernel in Fig. 1(right). If image f is constant, the result of this filter will be zero in every pixel. Show by calculation that the result for an image $f(x, y) = ay^2 + by + c$, is $-2a$ for each pixel with a, b, c constants. (Hint: just fill in the equation for discrete convolution for point (x, y), not some absolute location).

√ 8) (1.5 pt) **Thresholding.**
Consider the problem of thresholding a grey-level image f in which background and objects might vary in intensity.

a. (0.5pt) Describe the principle of global thresholding. Is it suitable for the problem above. Motivate your answer.

b. (0.5pt) Niblack's method computes a threshold using the equation

$$T = \mu_W(x, y) + k\sigma_W(x, y) \tag{1}$$

with $\mu_W(x, y)$ the mean grey value within window W centred around (x, y) and $\sigma_W(x, y)$ the standard deviation within the same window. Parameter k can be set by the user. How would you set k if you are looking for either bright or dark details? How does the absolute value of k affect the result? Motivate your answers.

b. (0.5pt) RATS uses the following statistic to compute a threshold in a region:

$$T = \frac{\sum_{(x,y)\in W} w(x, y) f(x, y)}{\sum_{(x,y)\in W} w(x, y)} \tag{2}$$

which is a weighted average of grey levels in the window W. What kind of operator is used as weight function w, and how do they deal with the presence of noise?

Some formulae:

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2})$$

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Discrete convolution by a kernel with support from $-W/2$ to $W/2$ in both dimensions:

$$(f * k)(x, y) = \sum_{u=-W/2}^{W/2} \sum_{v=-W/2}^{W/2} f(x-u, y-v)k(u, v)$$