SGN-41007 Pattern Recognition and Machine Learning
Exam 10.4.2018
Heikki Huttunen

- ▷ Use of calculator is allowed.
- ▷ Use of other materials is not allowed.
- ▷ The exam questions need not be returned after the exam.
- ▷ You may answer in English or Finnish.

1. Are the following statements true or false? No need to justify your answer, just T or F. Correct answer: 1 pts, wrong answer: $-\frac{1}{2}$ pts, no answer 0 pts.

   (a) Maximum likelihood estimators are unbiased.

   (b) Least squares estimator minimizes the squared distance between the data and the model.

   (c) The number of support vectors of a support vector machine equals the total number of samples.

   (d) Random forest has a linear decision boundary.

   (e) Maxpooling returns the maximum within each disjoint block of neighboring samples.

   (f) Stratified cross-validation resamples the training data such that all classes have equal number of samples.

2. The *Poisson distribution* is a discrete probability distribution that expresses the probability of a number of events $x = 0, 1, 2, \ldots$ occurring in a fixed period of time:

$$p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!},$$

   with parameter $\lambda > 0$ defining the shape of the density. In an experiment, N samples are measured to produce a Poisson distributed signal $x[n]$, $n = 0, 1, \ldots, N - 1$.

   (a) Find the maximum likelihood estimator of λ. (4p)

   (b) Is it unbiased? (2p)

3. A dataset consists of two classes, whose distributions are assumed Gaussian, and whose sample covariances and means are the following:

$$\boldsymbol{\mu}_0 = \begin{pmatrix} 3 \\ 1 \end{pmatrix} \qquad \boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}$$

$$\mathbf{C}_0 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \qquad \mathbf{C}_1 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

   Calculate the LDA projection vector $\mathbf{w}$.

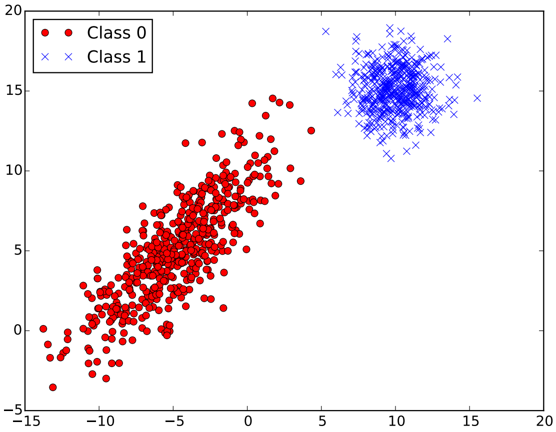

Figure 1: Training sample of question 3

	Prediction	True label
Sample 1	0.8	1
Sample 2	0.5	1
Sample 3	0.6	0
Sample 5	0.4	0
Sample 4	0.2	0

Table 1: Results on test data for question 5.

4. (6 pts) Consider the Keras model defined in Listing 1. Inputs are 128×128 color images from 10 categories.

   (a) Draw a diagram of the network.

   (b) Compute the number of parameters for each layer, and their total number over all layers.

5. A random forest classifier is trained on training data set and the `predict_proba` method is applied on the test data of five samples. The predictions and true labels are in Table 1. Draw the receiver operating characteristic curve. What is the Area Under Curve (AUC) score?

Listing 1: A CNN model defined in Keras

```
model = Sequential()

w, h = 3, 3
sh = (3, 128, 128)

model.add(Convolution2D(32, w, h, input_shape=sh, border_mode='same'))
model.add(MaxPooling2D(pool_size=(4, 4)))
model.add(Activation('relu'))

model.add(Convolution2D(48 w, h, border_mode='same'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Activation('relu'))

model.add(Convolution2D(64, w, h, border_mode='same'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Activation('relu'))

model.add(Flatten())
model.add(Dense(100))
model.add(Activation('relu'))

model.add(Dense(10, activation = 'softmax'))
```

Related Wikipedia pages

Inversion of 2 × 2 matrices [edit]

The *cofactor equation* listed above yields the following result for 2 × 2 matrices. Inversion of these matrices can be done as follows:[6]

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

ROC space [edit]

The contingency table can derive several evaluation "metrics" (see infobox). To draw a ROC curve, only the true positive rate (TPR) and false positive rate (FPR) are needed (as functions of some classifier parameter). The TPR defines how many correct positive results occur among all positive samples available during the test. FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test.

A ROC space is defined by FPR and TPR as *x* and *y* axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent to sensitivity and FPR is equal to 1 − specificity, the ROC graph is sometimes called the sensitivity vs (1 − specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space.

feature space by an SVM is an ellipse in the input space.

For degree-d polynomials, the polynomial kernel is defined as[2]

$$K(x,y) = (x^\mathsf{T} y + c)^d$$

where x and y are vectors in the *input space*, i.e. vectors of features computed from training or test samples and $c \geq 0$ is a free parameter trading off the influence of higher-order versus lower-order terms in the polynomial. When $c = 0$, the kernel is called homogeneous.[3] (A further generalized polykernel divides $x^\mathsf{T} y$ by a user-specified scalar parameter a.[4])

As a kernel, K corresponds to an inner product in a feature space based on some mapping φ:

$$K(x,y) = \langle \varphi(x), \varphi(y) \rangle$$

The nature of φ can be seen from an example. Let $d = 2$, so we get the special case of the quadratic kernel. After using the multinomial theorem (twice—the outermost application is the binomial theorem) and regrouping,

$$K(x,y) = \left(\sum_{i=1}^n x_i y_i + c\right)^2 = \sum_{i=1}^n \left(x_i^2\right)\left(y_i^2\right) + \sum_{i=2}^n \sum_{j=1}^{i-1} \left(\sqrt{2} x_i x_j\right)\left(\sqrt{2} y_i y_j\right) + \sum_{i=1}^n \left(\sqrt{2c}\, x_i\right)\left(\sqrt{2c}\, y_i\right) + c^2$$

From this it follows that the feature map is given by:

$$\varphi(x) = \langle x_n^2, \ldots, x_1^2, \sqrt{2} x_n x_{n-1}, \ldots, \sqrt{2} x_n x_1, \sqrt{2} x_{n-1} x_{n-2}, \ldots, \sqrt{2} x_{n-1} x_1, \ldots, \sqrt{2} x_2 x_1, \sqrt{2c}\, x_n, \ldots, \sqrt{2c}\, x_1, c \rangle$$