T-61.3050, Machine Learning: Basic Principles Exercise session 5/2015

The problems are divided into demonstration exercises and home assignments. The **deadline** for the home assignments is the **Monday, 23.11.2015, 10:15am**. Please, submit your answers via MyCourses. Alternatively, exercises can be returned on paper to a box (labeled with the course name) on the 3rd floor of the Computer Science Building. If you are submitting on paper, you still need to make a text submission within MyCourses. We prefer electronic submissions.

Note: Late submissions will not be graded. In such a case the system will not permit you to complete the submission.

Demonstration

1. Use the data set for Home assignment 2 of the exercise session 2 problem set (the data set of exchange rates as a function of time) to compare the AIC and BIC model selection criteria for polynomials of different degree. See slide 14 of Lecture 6 for definitions. (Hint: If you have already computed the values of the cost function in the training set for polynomials of different degrees then you don't have to do much.)

2. Consider the Naive Bayes Boolean classifier for multivariate binary data $\mathcal{X} = \{(r^t, \mathbf{x}^t)\}_{t\in\{1,\ldots,N\}}$, where $r^t \in \{0,1\}$ and $\mathbf{x}^t \in \{0,1\}^d$. Use parameters $p_r = P(r^t = 1)$ and $p_{ij} = P(x_j^t = 1 \mid r^t = i)$. See Section 5.7 in the course book.

   (a) Derive the maximum likelihood estimate of the parameters p_r and p_{ij} of the model given a data sample of size N.

   (b) Write down expression for classification probability $P(r = 1 \mid \mathbf{x})$ as a function of the parameters p_r, p_{ij}, and the new unclassified observation $\mathbf{x}$.

   (c) Show that the classification probability can be written in the sigmoid form

   $$P(r = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(w_0 + \mathbf{w}^{\mathrm{T}}\mathbf{x})}$$

   where $\mathbf{w} \in \mathbb{R}^d$ and $w_0 \in \mathbb{R}$ are functions of the parameters p_r and p_{ij}. (Further reading: See Mitchell's "Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression", available at `http://www.cs.cmu.edu/~tom/NewChapters.html`.)

Home assignments

1. Download the SPECT HEART data set from MyCourses. The data set concerns diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images and comes from the UCI Machine Learning Repository. In the data set, the SPECT image of each patient has been summarized as 22 binary features (partial diagnoses) extracted from the image.

   The task is to predict the class of a patient (the final diagnosis, normal or abnormal) in the test set based on the 22 image features. Note that in the provided data matrices, the first column of each row is the final diagnosis, the remaining 22 columns are the image features. Use the results of demonstration question 2 and the training set to find the maximum likelihood estimates of the parameters, as well as $\mathbf{w}$ and w_0. Compute classification errors on both training and test sets. Take a look at the parameters of the classifier. Can you say which features are important in differentiating normal and abnormal patients?

   (Hint: Sometimes the maximum likelihood estimate gives zero or unit probabilities, for example, when in a given class some variable has only one value. This creates problems (under- and over-flows) if in the test set this variable actually has several values. You can solve this by adding *a prior observation count* α to the estimated probabilities to avoid zero (or unit) probabilities. In practice, you should estimate the Bernoulli parameters with something like

   $$\hat{p}_{ij} = \frac{\alpha + \sum_t x_j^t r_i^t}{K\alpha + \sum_t r_i^t},$$

   where α is a positive constant (you can for example use $\alpha = 1$) and K is the number of classes.)

2. Implement the gradient descent algorithm to find parameters $\theta = (\mathbf{w}, w_0)$ for the logistic discrimination binary classifier, i.e. the algorithm should find the parameters that maximize

   $$\mathcal{L} = \sum_{t=1}^{N} \left(r^t \log y^t + (1 - r^t) \log(1 - y^t) \right)$$

   where $y^t = P(r^t = 1|\mathbf{x}^t) = \text{sigmoid}(\mathbf{w}^{\mathrm{T}}\mathbf{x}^t + w_0)$. The outline of the algorithm is given in the "Multivariate methods" lecture. You need to find partial derivatives $\partial\mathcal{L}/\partial\theta_i$ in order to implement the algorithm. Now use the "2-class Diabetes" "pima_indians_diabetes.csv" dataset (used in Exercise 4), find the solution for $\mathbf{w}$ and w_0, and compute training and test classification errors. Notice in Exercise 4 you classified the patients into the two different classes based on a single physical attribute only. But here in logistic regression, you are using all the attributes together to classify the patients. Compare the classification error you found before and now. Include the source code for gradient descent in the report, or send it as a separate file.

   (Hints: use small value for step size $\eta_s = 0.01$ and for the convergence criterion try comparing likelihoods between two successive iterations.)