Technische Universität München **SS 2016**
Fakultät für Informatik **Exercise Sheet 3**
Prof. N. Navab, Ph.D. **April 25, 2016**

Practical Course: Machine Learning in Medical Imaging

Linear Classifier and Evaluation Measurements

In this exercise, you can either implement all necessary functions yourself or use built-in functions of Scikit-learn. But please note that you should understand the algorithm behind a built-in function if you use it. If you have questions regarding the exercise, please contact tingying.peng@tum.de.

1 Logistic Regression

1.1 Likelihood-ratio Test

The likelihood-ratio test can be used to determine if considering additional features in the model results in an increased performance. The test statistic of the likelihood-ratio test D is defined as

$$D = -2 \log \left(\frac{\text{likelihood reduced model}}{\text{likelihood full model}} \right). \tag{1}$$

Under the null-hypothesis that the reduced model performs as well as the full model, D is χ^2 distributed with degrees of freedom (df) equal to the difference in the number of features considered. The p-value can be calculated using the upper incomplete gamma function:

$$\Gamma(a, z) = \frac{1}{\Gamma(a)} \int_z^\infty t^{a-1} e^{-t} dt \tag{2}$$

by calling `scipy.special.gammaincc(df/2, D/2)`. The resulting p-value gives an indication how likely it is that the result of the likelihood-ratio test arose just from chance. If the p-value is smaller than 0.05, the full model provides a significant benefit over the reduced model.

1.2 South African Heart Disease

The data set `SAheart` is a subset of the Coronary Risk-Factor Study (CORIS) baseline survey, carried out in three rural areas of the Western Cape, South Africa. The aim of the study was to establish the intensity of ischemic heart disease risk factors in that high-incidence region. The data represents white males between 15 and 64, and the response variable is the presence or absence of myocardial infarction (`chd`) at the time of the survey. The data consists of 160 cases, 302 controls and 9 features. The features are systolic blood pressure (`sbp`), cumulative tobacco in kg (`tobacco`), low density lipoprotein cholesterol (`ldl`), adiposity (`adiposity`), family history of heart disease (`famhist`), type-A behaviour (`typea`), obesity (`obesity`), current alcohol consumption (`alcohol`), age at onset (`age`) [1, 2].

Tasks

1. Create a logistic regression model for the `SAheart` data set, which takes a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ of samples and a vector $\mathbf{y} \in \{0; 1\}^n$ of outcomes for each sample. The function should return the *maximum likelihood estimate* (MLE) of the coefficients $\hat{\boldsymbol{w}}$ (including the intercept) and the log-likelihood of that model. All coefficients $\hat{\boldsymbol{w}}$ should be initialized with zeros. If you would like to use Sklearn built-in function `sklearn.linear_model.LogisticRegression` to obtain coefficients, you still need to compute the log-likelihood of the model. Moreover, make sure that you set the regularization strength to be small.

   a) Create a model that contains only the intercept (**null model**), i.e. no features are considered.

   b) Create multiple models each considering a single feature. Note that `famhist` is a categorical feature which has to be converted to numbers first.

   c) Create a function `likelihood_ratio_test` implementing the likelihood-ratio test which takes the log-likelihood of the full model and the reduced model (Section 1.1). Use this function to compare the *single feature models* to the *null model*. Which feature yields the most significant improvement over the null model?

   d) What do the estimated coefficients tell with respect to the odds of suffering from myocardial infarction? Make sure you consider the p-value of the likelihood-ratio test as well.

   e) Create a model which considers multiple features by starting with the null model and adding one additional feature at a time. To determine which feature to add, use the p-value as returned by the likelihood-ratio test. Extended models with one additional feature, where the p-value is greater than 0.05, should not be considered. In each step choose the model with the smallest p-value. Continue until all features have been selected or the model cannot be improved significantly any more. Print all selected features.

   f) L1 (lasso) regularization can also be used for feature selection. Consider a full model with all 10 features as input, penalized with the L1 norm of coefficients (try regularization parameter `C` in the range of $0.01 - 0.1$). Features with an non-zero coefficient are important for the classification. Compare the Lasso-selected features to the features selected by p-values. Please note Lasso-feature selection requires a standardization of features that each feature has a zero mean and a unit standard derivation (e.g.using Sklearn built-in function `sklearn.preprocessing.scale`)

2 ROC and Precision-Recall Curve

Now assume that we can obtain multiple confusion matrices because our *binary* classifier is able to assign each prediction a probability or score. Given a threshold t all predictions with probability smaller than t are classified as negative and positive otherwise. Gradually increasing this threshold from 0 to 1, we can construct one confusion matrix for each threshold.

Tasks

1. Create a function `threshold_confusion_matrix` that expects a vector containing the ground truth and a vector of containing the predicted probabilities for each sample. For each **unique** threshold the function should return a 2×2 confusion matrix.
2. Based on the list of confusion matrices obtained by `threshold_confusion_matrix` you can easily derive all the performance measures you already implemented. Construct a ROC and precision-recall curve for the different logistic regression models you created for the *South African Heart Disease* data set. Calculate the area under the curve. Which model performs best?

3 Cross Validation

Previously, we assessed the performance of our models merely on the training error, which results in bad estimates of the classifier's true performance. Hence, we want to train and test the classifier on disjoint sets by cross-validation.

Tasks

1. In the full logistic regression model (using all features), we can add a L2 based regularization term. Use cross-validation to find out the optimal regularization strength. Divide the datasets into 3 folds in the cross-validation.
2. Change the L2 based regularization term into L1 (Lasso) regularization and repeat the task.

References

[1] J. Rousseauw, J. du Plessis, A. Benade, P. Jordaan, J. Kotze, and J Ferreira. Coronary risk factor screening in three rural communities. *South African Medical Journal*, 64:430–436, 1983.

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*, pages 122–124. Springer, second edition, 2009.

[3] John Verzani. *Using R for Introductory Statistics*, page 296. Chapman and Hall, 2004.