Evaluation methods

Tasks (Lab 5):

1. Prepare two datasets:
   (a) Choose one real dataset corresponding to binary classification problem.
   (b) Create artificial dataset in the following way:

   $$y_i \sim Bern(p_i),$$

   where

   $$p_i = \frac{1}{1+\exp[-(\alpha+\beta^T x_i)]},$$

   for $i = 1, \ldots, n$, $x_i \sim N(0, I)$ and $\beta = (b, b, b, b, b, 0, \ldots, 0)$ (there are 5 relevant variables and k irrelevant variables). We treat α, b, k and n as parameters which will vary in simulations.

2. Fit two simple classification models: logistic regression and classification tree.

3. Assess its performance, for one chosen setting of parameters (e.g. $n = 1000$, $b = 1$, $k = 20$). Estimate classification error using different schemes.
   (a) Refitting, i.e. use the whole dataset for both training and testing.
   (b) 10-fold cross-validation.
   (c) Boostrap method.
   (d) Boostrap 0.632.

4. Draw ROC curve and precision-recall curve using the whole data. Try different values of parameters for artificial dataset $n = 100, 1000$, $b = 0.5, 1$, $k = 5, 50$.

5. Split data into training (50%) and testing (50%) sets. Fit the model using training and calculate accuracy and balanced accuracy using testing data using classification rule

   $$P(y_i = 1|x_i) > t,$$

   where t is a threshold. Draw plots showing how accuracy and balanced accuracy depend on t. Observe that $t = 0.5$ is the optimal threshold for precision, while $t = p(y = 1)$ is the optimal threshold for balanced accuracy.