Homework 2
due November 17, 16-00.

Skoltech
Machine learning course
November-December 2015

Recommendations: all solutions should be short, mathematically strict (unless qualitative explanation is needed), precise with respect to the stated question and clearly written.

Notification: lecture+seminar on November 17 will last longer than usual: from 16-00 till 19-45.

1. Consider classification with 1-nearest neighbour:
   (a) Prove that the decision boundary for 2 training objects of different classes will be linear.
   (b) Explain why decision boundaries, separating classes in 1 nearest neighbour classifier will be (possibly disjoint) piecewise linear curves for N training objects and C classes.

2. We studied Bayes minimum cost decision rule for the case of 2 classes ω_1 and ω_2 when costs of misclassification are $cost(\widehat{\omega}_2, \omega_1) = \lambda_1$ and $cost(\widehat{\omega}_1, \omega_2) = \lambda_2$ ($\widehat{\omega}$ stands for prediction of actual value ω).
   (a) Write down Bayes minimum cost decision rule for the case of C classes: $\omega_1, \omega_2, ...\omega_C$ with costs of misclassification $cost(\widehat{\omega}_k, \omega_i) = \begin{cases} 0, & k = i \\ \lambda_i, & k \neq i \end{cases}$
   (b) Prove that Bayes minimum cost decision rule reduces to predicting most probable class

$$\widehat{\omega}(x) = \arg\max_{\omega} p(\omega|x)$$

   when $\lambda_1 = \lambda_2 = ... = \lambda_C = \lambda$.

3. Prove that the complexity (number of elementary mathematical operations such as +,-,*,/ for scalars and boolean condition checks) for binary decision tree training from training set with N objects, having D features each:
   (a) does not exceed $O\left(DN^2 \log_2 N\right)$
   (b) may be reduced to $O\left(DN \left(\log_2 N\right)^2\right)$ if we use economic class probabilities recalculation within each node. Describe explicitly what that economic recalculation should be?

4. Consider binary classification trees.
   (a) Explain qualitively why decision trees with checks of individual features

$$\text{for node } t\text{:} \quad \begin{cases} x^{i(t)} \leq \gamma_t & \text{follow left child of } t \\ x^{i(t)} > \gamma_t & \text{follow right child of } t \end{cases}$$

   may be inaccurate when actual (true) class separating boundary is not parallel to axes of the feature space.
   (b) Suggest an idea of possible algorithm of binary decision tree construction where within each node t the split is based on whether linear combination of all features is greater or less than threshold:

$$\text{for node } t\text{:} \quad \begin{cases} \alpha_t^T x \leq \gamma_t & \text{follow left child of } t \\ \alpha_t^T x > \gamma_t & \text{follow right child of } t \end{cases}$$

   $\alpha_t \in \mathbb{R}^D$, $\gamma \in \mathbb{R}$. The algorithm should outline the idea how α_t and γ_t may be found for each t and state some possible stopping criterion setting the current node to the internal/terminal node of the tree.