Lehrstuhl für Informatik VI **Thursday, December 6th, 2018**
Rheinisch–Westfälische Technische Hochschule Aachen
Prof. Dr.–Ing. H. Ney **{kim,kitza,gao}@cs.rwth-aachen.de**

7. Exercise Sheet **Statistical Classification and Machine Learning**

Solutions to the problems indicated by (*...) must be submitted by **08:00 of Thursday, December 13th, 2018** via L2P. Please form a group of up to **three** students! (Group Workspace in L2P)

1. **Empirical Distribution and Classification Error Rate**

   For this task, you are given discrete training data (`data.txt`) with observation $x \in \{0, 1, ..., 9\}^2$ and class label $c \in \{0, 1, ..., 4\}$. Each line is in the format of:

   $$[x_1,\ x_2]\ c$$

   (a) Calculate the following empirical probabilities: (* 2P)
   - $p(c = 0)$
   - $p(x_1 = 0, x_2 = 0)$
   - $p(x_1 = 0, x_2 = 0|c = 0)$
   - $p(c = 0|x_1 = 0, x_2 = 0)$

   Please write your answers with fraction, e.g. $\frac{\text{count}_2}{\text{count}_1}$.

   (b) Give the formula of the error rate of a general decision rule. (* 1P)

   (c) Calculate the error rate of the following decision rules: (* 6P)
   - $x \mapsto r(x) = \underset{c}{\text{argmax}}\ \{p(c|x_1)\}$
   - $x \mapsto r(x) = \underset{c}{\text{argmax}}\ \{p(c|x_2)\}$
   - $x \mapsto r(x) = \underset{c}{\text{argmax}}\ \{p(c|x_1, x_2)\}$
   - $x \mapsto r(x) = \underset{c}{\text{argmax}}\ \{p(x_1, x_2|c)\}$

   If you see a division by zero, use prior probability $p(c)$ instead.

   (d) Interpret the results of (c). (* 1P)

2. **Error Bounds**

   Let us denote a true (posterior) distribution by $pr(c|x)$ and a model by $q(x, c)$ over input observations x and class label c. For a given input x, we define two decision rules—one with the true distribution and the other with the model:

   $$x \mapsto c_*(x) = \underset{c}{\text{argmax}}\{pr(c|x)\}$$
   $$x \mapsto c_q(x) = \underset{c}{\text{argmax}}\{q(x, c)\}$$

   (a) Define a classification error count $e_q(x, c)$ of the model q for a given input x and an assumed true class c. Derive a model-based local classification error $E_q(x)$, which is an expectation of the error count over all possible classes. (* 1P)

(b) After some derivations, the difference between the model-based error and the Bayes error (of the true distribution) for a given x is bounded by: (* 3P)

$$\begin{aligned} E_q(x) - E_*(x) &= ... \\ &\leq |pr(c_*(x)|x) - q(x, c_*(x))| + |pr(c_q(x)|x) - q(x, c_q(x))| \end{aligned}$$

From this formula, show that it can further induce the following three bounds:

$$E_q(x) - E_*(x) \leq \sum_c |pr(c|x) - q(x, c)| \qquad (l_1 \text{ bound})$$

$$E_q(x) - E_*(x) \leq 2 \cdot \max_c |pr(c|x) - q(x, c)| \qquad (l_\infty \text{ bound})$$

$$E_q(x) - E_*(x) \leq 2 \cdot \sqrt{\sum_c [pr(c|x) - q(x, c)]^2} \qquad (l_2 \text{ bound})$$

(c) Now we consider the global error bounds for all possible input x. How is the l_1 bound of (b) changed? What will become of the square of the global bound? (* 1P)

(d) Assume that our model q is normalized over c and we denote it by $q(c|x)$. The quadratic bound of (c) develops into: (* 2P)

$$\begin{aligned} (E_q(x) - E_*(x))^2 &\leq ... \\ &\leq 2 \cdot \sum_x pr(x) \sum_c pr(c|x) \log \frac{pr(c|x)}{q(c|x)} \end{aligned}$$

Show that the optimal model $\hat{q}(c|x)$ that minimizes this upper bound is indeed the true distribution $pr(c|x)$, using the divergence inequality:

$$\sum_c p_c \log \frac{p_c}{q_c} \geq 0$$

for two distributions p_c and q_c over a random variable c.

(e) Convert the error bound of (d) to the cross entropy training criterion. To use this criterion in practice, how should we change the assumption about the distributions? (* 2P)

(f) Explain why minimizing the error bound is effective in building a good classification model. (* 1P)