Introduction to Machine Learning in R  
(NPFL054)

Easy HW – SVM  
Contact: Barbora Hladká (hladka@ufal.mff.cuni.cz)

Data

- You will use data set Caravan, which is a part of ISLR package available in R. This is the same dataset you experiment with in Homework #2.  
  \`\`dim(Caravan) = 5,822 x 86\`

Get the data for the experiments

- Split the data in two parts
  - development test set \`\`Dtest\`\` – randomly selected 1,000 examples
  - development working data set \`\`Dtrain\`\` – the remaining examples

- Randomly divide \`\`Dtrain\`\` into 8 folds (\`\`D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8\`\`) to run cross-validation. Since the positive and negative examples are highly unbalanced, you should make the division carefully. You should always keep the identical number of positive examples in all your folds.

Questions

1. Address the task of Purchase prediction using SVM with Radial basis kernel. Use \`\`Dtrain\`\` as a development data to run 8-fold cross validation with the folds \`\`D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8\`\`. Use the function \`\`svm()\`\` with \texttt{kernel=“radial”}. Report cross-validation error rates for various values of \texttt{gamma} and \texttt{cost}.

2. Address the task of Purchase prediction using SVM with polynomial kernel. Use \`\`Dtrain\`\` as a development data to run 8-fold cross validation with the folds \`\`D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8\`\`. Use the function \`\`svm()\`\` with \texttt{kernel=“polynomial”} and \texttt{gamma=1}. Report cross-validation error rates for various values of \texttt{cost} and \texttt{degree}.

Presentation

- Create a 5 min presentation.
- Present your answers. If you want to highlight something in your R code, please do it.
- Explain your answers clearly so that your audience understands your method well.