Introduction to Machine Learning
NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká **Martin Holub**

{Hladka | Holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Programming questions

- NLI task, `train` and `test` data sets
  - Train SVM with Linear kernel on `train` and do prediction on `test`
  - Train SVM with Radial Basis kernel on a (non)scaled subset of `train` and do prediction on `test`

- College data set
  - `https://cran.r-project.org/web/packages/ISLR/ISLR.pdf`
  - Train Logistic regression classifier
  - Train Decision tree classifier
  - Evaluate both classifiers using ROC curve and AUC measure

Native Language Identification (NLI)

Identifying the native language (L1) of a writer based on a sample of their writing in a second language (L2)

Our data

- **L1s**: Arabic (ARA), Chinese (ZHO), French(FRA), German (DEU) Hindi (HIN), Italian (ITA), Japanese (JPN), Korean (KOR), Spanish (SPA), Telugu (TEL), Turkish (TUR)
- **L2**: English
- **Real-world objects**: For each L1, 1,000 texts in L2 from The ETS Corpus of Non-Native Written English (former TOEFL11), i.e. $Train \cup DevTest$
- **Target class:** L1

More detailed info is available at the course website.

NLI
Features used

96 numerical features = relative character frequencies

Example

"Finally having people with many academic broad know"

```
  <SPACE>          a          b          c          d          e
0.17073171 0.14634146 0.02439024 0.04878049 0.04878049 0.07317073
         m          n          o          F          g          h
0.04878049 0.09756098 0.07317073 0.02439024 0.02439024 0.04878049
         i          k          l          p          r          t
0.09756098 0.02439024 0.07317073 0.04878049 0.02439024 0.02439024
         v          w          y
0.02439024 0.04878049 0.04878049
```

Support Vector Machines in R

- `library(e1071)`, but there are also other libraries (**kernlab**, **shogun** ...)
- training: function `svm()`
- prediction: function `predict()`
- `svm()` can work in both classification and regression mode
- if target attribute (response variable) is categorical (factor) the engine switches to classification

SVM in R

```
model = svm(formula, data=, kernel=, cost=, cross=, ...)
```

- `?svm`
- `kernel` defines the kernel used in training and prediction. The options are: linear, polynomial, radial basis and sigmoid, default = radial
- `cost` – cost of constraint violation (default: 1)
- `cross` – optional, with the value k the k-fold cross-validation is performed

SVM kernels in `e1071`

Kernel name	Formula	Learning parameters and their default values
linear	$\mathbf{x}_i \cdot \mathbf{x}_j$	
polynomial	$(\gamma \mathbf{x}_i \cdot \mathbf{x}_j + c)^d$	γ, `gamma`=1/(data dimension) c, `coef0`=0 d, degree=3
radial	$\exp(-\gamma(\|\|\mathbf{x}_i - \mathbf{x}_j\|\|^2))$	γ, `gamma`=1
sigmoid	$\tanh(\gamma \mathbf{x}_i \cdot \mathbf{x}_j + c)$	γ, `gamma`=1/(data dimension) c, `coef0`=0

SVM – non-linear kernel functions

- polynomial kernel
  - smaller degree can generalize better
  - higher degree can fit (only) training data better

- radial basis
  - very robust
  - you should try and use it when polynomial kernel is weak to fit your data

SVM Parameter tuning with `tune.svm`

- **SVM is a more complicated method in comparison with the previous and usually requires parameter tuning!**
- parameter tuning can take a very long time on big data, use a reasonably smaller part is often recommended

```
> model.tune= tune.svm(class ~ ., data=train.small,
                     kernel = "radial",
                     gamma = c(0.001, 0.005, 0.01, 0.015, 0.02),
                     cost = c(0.5, 1, 5,  10))
> model.tune
Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:
 gamma cost
 0.01   1

- best performance: 0.739
```

Built-in cross-validation

K-fold cross-validation

- parameter cross

```
> model.best <- svm(class ~ ., train.small,
                                  kernel = "radial",
                                  gamma = 0.01,
                                  cost = 10,
                                  cross = 10)
> model.best$accuracies
[1] 26.81 30.90 36.36 28.63 38.18 28.18 37.72 35.90 34.09 30.90
> model.best$tot.accuracy
[1] 32.77
> prediction.best <- predict(model.best, test, type="class")
> mean(prediction.best==test$class)
[1] 33.36
```

Class weighting

- `class.weights` parameter
  In case of asymmetric class sizes you may want to avoid possibly overproportional influence of bigger classes. Weights may be specified in a vector with named components, like
  `m <- svm(x, y, class.weights = c(A = 0.3, B = 0.7))`

General hints on practical use of `svm()`

- Note that SVMs may be very sensible to the proper choice of parameters, so always check a range of parameter combinations, at least on a reasonable subset of your data.
- Be careful with large datasets as training times may increase rather fast.
- C-classification with the `RBF` kernel (default) can often be a good choice because of its good general performance and the few number of parameters (only two: `cost` and `gamma`).
- When you use C-classification with the `RBF` kernel: try small and large values for `cost` first, then decide which are better for the data by cross-validation, and finally try several `gamma` values for the better `cost`.

ROC curve

Receiver Operator Characteristics curve: FPR vs. TPR

- $[1, 1]$ – all of the known positives were classified correctly, all of the known negatives were classified incorrectly
- e.g. $[0.75, 1]$ is to the left of the diagonal: the proportion of correctly classified known positives is greater than the proportion of incorrectly classified known negatives
- e.g. $[0, 0.75]$ is on y axis: 75% of known positives and 100% of known negatives were classified correctly
- $[0, 0]$ – all of the known positives were classified incorrectly, all of the know negatives were classified incorrectly