IST Austria: Statistical Machine Learning 2015/16

Christoph Lampert <chl@ist.ac.at>
TA: Alex Zimin <azimin@ist.ac.at>

Exercise Sheet 4/6

1 Properties of the VC-Dimension

a) What is the VC-dimension of the hypothesis class of all classifiers in $\mathbb{R}$ with the property that their positive region is a finite interval, or that their negative region is a finite interval?

b) What is the VC-dimension of the hypothesis class of all classifiers in $\mathbb{R}^2$ with the property that either their positive or negative region is a rectangle?

c) What is the VC-dimension of the hypothesis class of all classifiers in $\mathbb{R}^d$ with the property that either their positive or negative region is a ball with arbitrary center and radius?

d) Prove or disprove: if $\mathcal{H}$ an $\mathcal{H}'$ are hypothesis classes and $\mathcal{H} \subset \mathcal{H}'$, then $\mathrm{VCdim}(\mathcal{H}) \leq \mathrm{VCdim}(\mathcal{H}')$.

e) Let $\mathcal{H}$ an $\mathcal{H}'$ be hypothesis classes with finite VC-dimensions. Prove or disprove:

$$\mathrm{VCdim}(\mathcal{H} \cup \mathcal{H}') \leq \mathrm{VCdim}(\mathcal{H}) + \mathrm{VCdim}(\mathcal{H}') + 1.$$

f) Construct a *finite* $\mathcal{H}$ with $|\mathcal{H}| = 2^K$ elements, but $\mathrm{VCdim}(\mathcal{H}) \neq K$.

2 PAC Learning

Give an example why "PAC learning" isn't just "PC learning": show that even for a hypothesis class with just 2 elements no finite m_0 can guarantee that the hypothesis with 0 generalization error will be found with high probability, if arbitrary distributions are allowed.

3 No Free Lunch

The No Free Lunch theorem proves unrefutably that every learning algorithm will fail on some problems that another algorithm will easily solve. Discuss: why do we still do research on machine learning, and why do we still consider some algorithms "better" than others?

4 Useful Inequalities

Statistics and probability theory provide several inequalities that are useful when analyzing error probabilities.
1) **Markov's inequality**: let Z be a random variable that takes only non-negative values and has finite expected value, $\mathbb{E}[Z]$. Then it holds

$$\forall a > 0 : \ \Pr\{ Z \geq a \} \leq \frac{\mathbb{E}[Z]}{a}$$

a) Is it possible that more than half of the population have a salary more than twice the average salary?
b) Is it possible that more than 90% of the population have a salary less than 10% of the average salary?
c) For any $a > 0$, find a distribution for which Markov's inequality is *tight*, i.e. the inequality is an equality.
d) Give an example that the inequality can be violated if Z can take negative values.

2) Prove the following lemma: Let Z be a random variable and $\theta \in \mathbb{R}$ be a scalar. Assume that there exists $a > 0$ such that for all $t \geq 0$ we have

$$\Pr\{|Z - \theta| > t\} \leq 2e^{-t^2/a^2}.$$

Then,

$$\mathbb{E}[\,|Z - \theta|\,] \leq 4a.$$

Hint: for $i = 0, 1, 2, \ldots$, define $t_i = ai$. First, check that $\mathbb{E}[\,|X - \theta|] \leq \sum_{i=1}^{\infty} t_i \mathbb{P}[|X - \theta| > t_{i-1}]$. Assume the assumption of the lemma to show that $\mathbb{E}[\,|X - \theta|] \leq 2a \sum_{i=1}^{\infty} t_i i e^{(i-1)^2}$. Then prove that the infinite sum is bounded by 2.

5 Practical Experiments VI

a) Construct a synthetic learning problem in $\mathbb{R}^2$ with classes that are linearly separable with a margin and for which you can compute the generalization error of any linear classifier (ideally analytically, but otherwise approximately numerically). For different amounts of training data, train a linear classifier with zero training loss (e.g.a perceptron or hard-margin SVM) and plot the generalization error. How does it scale with respect to the amount of training data? How close is the error to the upper bound from the value in the generalization bound?
b) Construct an synthetic learning problem in $\mathbb{R}^2$ for which classes cannot be separated perfectly. For different amounts of training data, train a linear classifier and plot the difference between training error and generalization error. How does this quantity scale with respect to the amount of training data? How close is its value to the upper bound from the generalization bound?
c) In the setting of b), try to construct or compute a classifier with as big difference between training and test error as possible. Can you find a training set for which the difference exceeds the bound? How often does this happen for 100 randomly sampled training sets?

6 Practical Experiments VII

Download the `census` dataset from the homepage. It consists of 45222 samples of two labels, split into *train*, *validation* and *test* part. The first column in the text file contains the label as well as flag `'v'` for *validation* and `'t'` for *test* samples.
Try all your existing classifier implementations on the data. If training time is too high, subsample the training set until training becomes possible.

1) Use the validation set for model selection, and report results on the test set.

2) For different random subsets of validation data (random 1%, random 10%, all 100% of available data), make a plot of the model complexity (e.g. C or λ parameter for SVM, depth of the decision tree, etc.) versus a) the validation error, and b) the value of a suitable bound, if there is one. Do they have the same minimum? Finally, also plot the error on the (full) test set. Discuss the results.