IST Austria: Statistical Machine Learning 2018/19
Christoph Lampert <chl@ist.ac.at>
TAs: Nikola Konstantinov <nkonstan@ist.ac.at>,
Mary Phuong <bphuong@ist.ac.at>
Amelie Royer <aroyer@ist.ac.at>
Exercise Sheet 4/6 (due date 5/11/2018)

1 Implicit and explicit bias term in maximum margin classification

In the lecture, we used the *augmentation* trick, $x \mapsto \tilde{x} = (x, 1)$, turn an affine decision function $\tilde{f}(x) = \langle w, x \rangle + b$ into a linear function $f(x) = \langle \tilde{w}, \tilde{x} \rangle$, with $\tilde{w} = (b, w)$. For SVM training, both formulations are similar, but not exactly equivalent. In this exercise, we study the differences.
a) Write down the learning problem for the maximum-margin classifier for the augmented data $\tilde{x}^i$ and weight vector $\tilde{w}$, and convert it to a form using x^i, w, and b. What's the difference to an SVM with explicit bias term?
b) Which of both versions make more sense (geometrically?)
c) In practice, one sometimes augments $\tilde{x} = (x, c)$ with a large $c > 0$ (e.g. 1000) instead of $c = 1$. Why?
d) Construct a two-class dataset in $\mathbb{R}^2$ for which the maximum margin classifier with explucit bias term and the one with implicit bias term (by augmentation) are very different. How close to orthogonal can you make them?

We now know three variants of the maximum margin classifier: 1) linear with no bias term, 2) affine with explicit bias b, 3) affine with implicit bias (by augmentation). Answer the following question for each of them:
e) What happens if you take the x-part of the training set and scale every vector by a fixed constant (isotropic scaling). Can you express the optimal weight vector (and bias, if present) for the scaled data in terms of the value(s) for the original data?

2 Error Bounds

In the lecture, we saw that the σ-rules for Gaussians do not necessarily hold for other distributions.
a) Find a distribution on the interval $[-1, 1]$ for which the inequality

$$\Pr\{|Z - \mathbb{E}[Z]| \geq 2\sigma\} \leq 0.05,$$

where σ^2 is the distribution's variance, is *maximally violated*.

In the lecture we saw **Chebyshev's inequality**:

$$\forall a \geq 0: \quad \mathbb{P}[|Z - \mathbb{E}[Z]| \geq a] \leq \frac{\mathrm{Var}[Z]}{a^2},$$

and as a concrete example we looked at the distribution $p(-1) = \frac{1}{10}, p(0) = \frac{4}{5}, p(1) = \frac{1}{10}$.
b) Show: there is an $a \geq 0$ for this distribution that makes the inequality *tight*, i.e.$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq a] = \frac{\mathrm{Var}[Z]}{a^2}$.

c) For arbitrary $a \geq 0$, find a distribution such that the Chebyshev inequality is tight with this value of a.

d) Show: every *linear transformation* of the above distribtion (i.e. $Z \mapsto cZ + b$), also has an a where it is tight.

e) Bonus puzzle: show that every distribution on $\mathbb{R}$ that has a point, a, where Chebyshev's inequality is tight has a form $p(-1) = \epsilon$, $p(0) = 1 - 2\epsilon$, $p(1) = \epsilon$ for some ϵ, or a linear transformation of it.

3 Estimators.

Let $S_n = \{z^1, z^2, \ldots, z^n\}$ be independent samples from $\mathcal{N}(x; \mu, \sigma^2)$. For the following estimators of σ^2, determine i) their *bias*, ii) their *variance*, iii) if they are *consistent*.

- $\hat{E}(S_n) = 1$
- $\hat{E}(S_n) = \frac{1}{n}\sum_{i=1}^n z_i^2 - (\frac{1}{n}\sum_{i=1}^n z_i)^2$
- $\hat{E}(S_n) = \frac{1}{n}\sum_{i=1}^n (z_i - \hat{\mu})^2$ with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n z_i$
- $\hat{E}(S_n) = \sum_{i=1}^n (z_i - \hat{\mu})^2$
- $\hat{E}(S_n) = \frac{1}{n-1}\sum_{i=1}^n z_i^2 - (\frac{1}{n-1}\sum_{i=1}^n z_i)^2$
- $\hat{E}(S_n) = \frac{1}{n-1}\sum_{i=1}^n (z_i - \hat{\mu})^2$ with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n z_i$

4 Practical Experiments VI

The so-called "Least Squares SVM" classifier is defined by the following optimization problem

$$\min_w \quad \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^n (w^\top x^i - y^i)^2$$

where $\lambda \geq 0$ is the regularization parameter. The LS-SVM has a closed-form solution for w:

$$w = (X^\top X + \lambda \,\mathrm{Id}_{d\times d})^{-1} X^\top Y$$

where $X \in \mathbb{R}^{n\times d}$ is the data matrix, i.e.the i-th row of X is $x^i \in \mathbb{R}^d$, and $Y \in \mathbb{R}^n$ is the vector of labels, i.e.the i-th entry of Y is $y^i \in \{\pm 1\}$.

- Download the *"Xtrain.txt"* and *"Ytrain.txt"* files from the website. Split the data randomly into two parts of (approximately) equal size: $Xtrn/Ytrn$, which we will use for learning the model, and $Xval/Yval$, to use as validation data.
- Implement the LS-SVM and learn a model without regularization ($\lambda = 0$) on *Xtrn/Ytrn*.
- Compute the 0/1-error on the training data and on the validation data. *What do you observe? Overfitting, underfitting or neither?*
- repeat the above steps of training and evaluation for $\lambda \in A$ for $A = \{2^{-20}, 2^{-19}, \ldots, 2^{20}\}$
- plot a graph with λ on the x-axis (in logarithmic units) and training and validation error on the y axis.
- Based on the results, pick a suitable regularization constant, λ, and use it to train a model on the combination of $Xtrn/Ytrn$ and $Xval/Yval$.
- Now (and only now, not before!) download the test data, *"Xtest.txt"* and *"Ytest.txt"*. There's no direct link, just change the filenames in the URLs of the training data. Evaluate the final classifier on this and report the result.

5 Practical Experiments VII

- Implement a *linear support vector machine (SVM)* with training by Stochastic Coordinate Dual Ascent.
- Implement a *linear SVM* with training by Stochastic Gradient Descent (e.g. adapting your code from Exercise sheet 3).
- What error rates do both methods achieve on the datasets from the previous exercise?
- Plot the number of steps against the values of the objective function and the error rate for the stochastic gradient method for different stepsizes, η, and for SCDA. What do you observe?