Statistical Machine Learning
Christoph Lampert <chl@ist.ac.at>
TA: Asya Pentina <apentina@ist.ac.at>
Mailing List: <machinelearning_s14@lists.ist.ac.at>
Exercise Sheet 3/5

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?

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 questions for each of them:
d) 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?
e) What happens if you take the x-part of the training set and shift every vector by a fixed vector (translation). Can you express the optimal weight vector (and bias, if present) for the translated data in terms of the value(s) for the original data?

2 Error Bounds

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 $p(-1) = \frac{1}{10}, p(0) = \frac{4}{5}, p(1) = \frac{1}{10}$.

a) Show: there is an $a \geq 0$ for this distribution that makes the inequality *tight*, i.e.$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq a] = \frac{\mathrm{Var}[Z]}{a^2}$.

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

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

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

3 Missing Proofs.

Prove **Hoeffding's Lemma**:
Let Z be a random variable that takes values in $[a, b]$ and fulfills $\mathbb{E}[Z] = 0$. Then, for every $\lambda > 0$,

$$\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2 (b-a)^2}{8}}.$$

Hints:
1) First show that $\mathbb{E}[e^{\lambda X}] \leq \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}$ by using the definition of convexity on the (convex) function $\exp(x)$.
2) For $h = \lambda(b - a)$, find $L(h)$ such that $\frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b} = e^{L(h)}$.
3) Show $L(h) \leq \frac{h^2}{8}$ (which completes the proof).

4 Empirical Risk Minimization

a) Prove: the hypothesis class of *axis-aligned rectangles* in $\mathbb{R}^2$ can be learned by the ERM algorithm, when the true labeling function is also an axis-aligned rectangle (when ERM is not unique, choose the smallest rectangle). Hint: use the same arguments as we did for thresholds on each coordinate, but beware that now the data distribution is not necessarily uniform, and the rectangles can lie everywhere in the plane, not just within $[0, 1]$. To merge the statements about the four coordinates, use a union-bound.
b) *(this can be done without doing a) first)*:
PAC learning can be succesful without returning a hypothesis that is "close" to the labeling function f itself. Prove: The coordinates, $[a, b, c, d]$, returned by the ERM algorithm can be arbitrarily far from the "true" coordinates $[a_f, b_f, c_f, d_f]$ that define f, in the sense that for any $\rho > 0$, possibly $|a - a_f| \geq \rho$, $|b - b_f| \geq \rho$, etc. Hint: for any $\rho > 0$, construct a suitable data distribution.

5 Practical Experiments IV

In the lecture we saw **Hoeffding's inequality**: For $Z_1, \ldots, Z_m \overset{i.i.d.}{\sim} p$ and $\mu = \mathbb{E}[Z_i]$,

$$\mathbb{P}[\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon] \leq 2e^{-m\frac{\epsilon^2}{(b-a)^2}},$$

For a Bernoulli variables Z_i write a program that for fixed μ, ϵ and m:

- estimates the left hand side from 100 random experiments,
- computes the right hand side value,
- plots the values in a graph with sample size on the x-axis and the bound or estimate on the y-axis.

Do this for a) $\mu = 0.5$, $\epsilon = 0.25$, $m = 1, 5, 10, 15, \ldots, 50$, b) $\mu = 0.1$, $\epsilon = 0.25$, $m = 1, 5, 10, 15, \ldots, 50$, c) $\mu = 0.5$, $\epsilon = 0.05$, $m = 10, 50, 100, 150, \ldots, 500$, d) $\mu = 0.1$, $\epsilon = 0.05$, $m = 10, 50, 100, 150, \ldots, 500$.

6 Practical Experiments V

- Pick one more training methods from the first sheet and implement it.
- In addition, implement a *linear support vector machine (SVM)* with training by Stochastic Coordinate Dual Ascent.
- What error rates do both methods achieve on the datasets from the previous sheet?
- 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?