IST Austria: Statistical Machine Learning 2020/21
Christoph Lampert <chl@ist.ac.at>
TAs: Alexandra Peste <alexandra.peste@ist.ac.at>,
Bernd Prach <bernd.prach@ist.ac.at>
Exercise Sheet 3/5 (due date 26/10/2020, 10:15am ← public holiday, no lecture!)
Please send your solutions via email to the TAs

1 Robustness of the Perceptron

Remember Perceptron training of Lecture 1 (deterministic with samples in fixed order). Look at the dataset with the following three points:

$$\mathcal{D} = \{\ (\begin{pmatrix} 2 \\ 1 \end{pmatrix}, +1), (\begin{pmatrix} 1 \\ -1 \end{pmatrix}, -1), (\begin{pmatrix} a \\ b \end{pmatrix}, +1)\} \subset \mathbb{R}^2 \times \{\pm 1\}.$$

- For any $0 < \rho \leq 1$, find values for a and b such that the Perceptron algorithm converges to a *correct* classifier with *robustness* ρ.
- What's the maximal robustness you can achieve for any choice of a and b?
- Can you find a situation (i.e. a,b) in which the classifier found by the perceptron algorithm has small robustness, but the max-margin classifier has much larger robustness?

Hint: it'll help to look at the Perceptron algorithm steps geometrically in 2D.

2 What's needed for a good test set?

In the lecture we saw that $\hat{\mathcal{R}}_{\text{tst}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(\,y_i, f(x_i))$ is an unbiased and consistent estimator of the true risk $\mathcal{R}(f) = \mathbb{E}_{(x,y)\sim p(x,y)}\ell(\,y, f(x)\,)$, if

a) each point (x_i, y_i) in the test set is sampled from the data distribution p

b) the points (x_i, y_i) are independent of each other

c) the predictor f is chosen independently of the test set

d) the loss function is bounded.

Show that each conditions a)–d) is necessary: for each condition construct a situation where the specific condition is violated but the others are fulfilled, and the $\hat{\mathcal{R}}_{\text{tst}}(f)$ is *not* an unbiased and consistent estimator of the true risk. In each case, state which property is violated, unbiasedness or consistency (or both).

Hint: *c*) entails that if (x_i, y_i) and (x_j, y_j) are independent of each other, then $\ell(y_i, f(x_i))$ and $\ell(y_j, f(x_j))$ are also independent random variables.

3 Properties of the Variance

a) Let Z be a random variable with finite mean and variance. Show that for any $a, b \in \mathbb{R}$ one has

$$\text{Var}[aZ + b] = a^2\text{Var}[Z]$$

b) Let Z be a random variable with finite expected value and variance. Show that

$$\text{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$$

c) Show that for independent random variables $Z_1, \dots, Z_n$ one has:

$$\mathrm{Var}\big[\sum_{i=1}^{n} Z_i\big] = \sum_{i=1}^{n} \mathrm{Var}[Z_i]$$

(tip: write the variance as a double sum over products; split the double sum into two: one where the factors in the product are independent and one where they are not; show that some terms vanish.)

d) Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function and let $\mathcal{D}_{\text{tst}} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ be a test set,sampled i.i.d. from the dataset distribution p. Let $f : \mathcal{X} \to \mathcal{Y}$ be a prediction function that was selected independently from $\mathcal{D}_{\text{tst}}$.

Prove the following statements:
1) for any bounded loss function, $\ell(y, \bar{y}) \in [0, M]$ for some $M > 0$, the following inequality holds

$$\mathrm{Var}\big[\hat{\mathcal{R}}_{\text{tst}}(f)\big] \leq \frac{M^2}{m},$$

2) if the loss function is 0/1-loss, then following *equality* holds

$$\mathrm{Var}\big[\hat{\mathcal{R}}_{\text{tst}}(f)\big] = \frac{\mu(1-\mu)}{m} \qquad \text{with } \mu = \mathcal{R}(f).$$

4 Bias of the Variance

Let $z_1, \dots, z_n$ be i.i.d. samples from a Gaussian distribution $\mathcal{G}(z; \mu, \sigma^2)$.

a) Here is two estimator of the mean:

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} z_i \qquad \text{and} \hat{\mu}_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} z_i$$

What's the *variance* of each of these estimators?

b) There's two popular estimator of the variance, σ^2:

$$\hat{\sigma}_n = \frac{1}{n} \sum_{i=1}^{n} (z_i - \hat{\mu})^2 \qquad \text{and} \qquad \hat{\sigma}_{n-1} = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \hat{\mu})^2 \qquad \text{for } \hat{\mu} \text{ as above.}$$

What the bias for each of these estimators? Which one has the bigger variance?

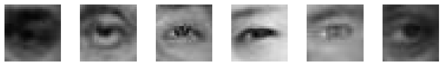
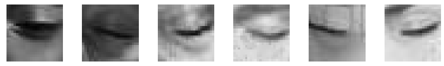
c) What if the samples are not from a Gaussian but from any real-valued random variable, Z, with expected value $\mathbb{E}[Z]$ and variance $\mathrm{Var}[Z]$. What changes in the answers to a) and b)?

5 Practical Experiments III

The goal of this exercise is observe overfitting and underfitting with different regularization strengths. As an example task we will classify images of open or closed eyes.

Setting

- **input**: x: images of eyes, grayscale, resolution 24×24 pixels

  examples: open eyes  closed eyes 

- **output**: $y = 1$: eye is open, $y = -1$: eye is closed.
- **model**: $g(x) = \operatorname{sign} f(x)$ for linear function, $f(x) = \langle w, x \rangle$. Model parameters: $w \in \mathbb{R}^d$
- **quality measure**: $\ell(y, f(x)) = [\![\operatorname{sign} f(x) \neq y]\!] = \begin{cases} 1 & \text{if } g(x) \neq y \\ 0 & \text{otherwise} \end{cases}$

Model Learning

- Load the data files *"XtrainIMG.txt"* and *"Ytrain.txt"*. They contain the training images as row vectors of pixel intensities and the corresponding ground truth annotation, respectively.
- Split the data randomly into two parts of (approximately) equal size: $\mathcal{D}_{\text{trn}}$, which we will use for learning the models, and $\mathcal{D}_{\text{eval}}$, which we will use to evaluate the models.
- For any regularization strength $\lambda \in \{2^{-15}, 2^{-14}, \dots, 2^{15}\}$ train a *least-squares classifier* on $\mathcal{D}_{\text{trn}}$ and evaluate its test error on $\mathcal{D}_{\text{eval}}$ (see below for details)
- plot a graph with λ on the x-axis (in logarithmic units) and training and validation error on the y axis.
- which regularization values result in overfitting, which ones in underfitting?
- repeat the above for two other classifiers: *logistic regression* and *soft-margin SVM* (with $C = \frac{1}{\lambda}$). How do the plots differ?

Please submit your code as well as the resulting plots.

Least-squares classifier

The least-squares classifier is a simple linear classifier that is popular (sometimes under the misleading name "Least-Squares SVM") because it has a closed-form solution:

For data $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, let $X \in \mathbb{R}^{n \times d}$ be the data matrix with columns $x_1, \dots, x_n$ and let $Y \in \mathbb{R}^n$ be the vector of label values. The least-squares classifier with regularization strength λ has the form

$$g(x) = \operatorname{sign}\langle w, x \rangle \quad \text{with} \quad w = (X^\top X + \lambda \operatorname{Id}_{d \times d})^{-1} X^\top Y$$

In python, it is available as `sklearn.linear_model.RidgeClassifier`.