EPFL, Autumn 2024

Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 1

Exercise 1.1 Classification from a discrete input space. We consider a multiclass classification problem with 3 classes $\{1, 2, 3\}$ for data with only a single discrete descriptor in $\mathcal{X} = \{1, 2, 3, 4\}$.

We assume that the joint probability distribution $\mathbb{P}(Y = y, X = x)$ with X taking values in $\mathcal{X}$ and Y taking values in $\mathcal{Y} = \{1, 2, 3\}$ is specified by the following table:

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	0,02	0,08	0,10
$X = 2$	0,05	0,40	0,15
$X = 3$	0,02	0,02	0,12
$X = 4$	0,02	0,01	0,01

(a) What is the target function f^* for the 0-1 loss?

Solution: For $0 - 1$ loss function, the risk is minimized when the target function assigns every x to the most likely class.

$$f^*(x) = \underset{y}{\operatorname{argmax}}\, \mathbb{P}[Y = y | X = x]$$

(b) What are the values of $f^*(x)$ for $x = 1, 2, 3, 4$.

Solution: Evaluating the above expression,

$$f^*(x) = \begin{cases} 3 & x = 1, 3 \\ 2 & x = 2 \\ 1 & x = 4 \end{cases}$$

(c) What is the value of the risk for the target function?

Solution: Evaluating the risk,

$$\begin{aligned} \mathcal{R}(f^*) &= \mathbb{E}[1_{\{f^*(X) \neq Y\}}] \\ &= \sum_{x=1}^{4} \sum_{y=1}^{3} 1_{\{f^*(x) \neq y\}} \mathbb{P}[X = x, Y = y] \\ &= \sum_{(x,y): f^*(x) \neq y} \mathbb{P}[X = x, Y = y] \\ &= 0,02 + 0,08 + 0,02 + 0,02 + 0,05 + 0,15 + 0,01 + 0,01 = 0,36 \end{aligned}$$

Exercise 1.2 Recap of linear models. Let $\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$, where $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ and $\mathbf{X}$ is a non-random full rank matrix of size $n \times p$. This setup contains the Gauss-Markov assumptions of a linear model.

(a) Derive the least squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$.

Solution: The residual sum of squares is given by $\mathrm{RSS}(\beta) = (\boldsymbol{y} - \mathbf{X}\beta)^\top(\boldsymbol{y} - \mathbf{X}\beta)$. Differentiating with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\beta).$$

Setting the first derivative to zero gives

$$\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\beta) = \mathbf{0}.$$

Since $\mathbf{X}$ has full column rank, $\mathbf{X}^\top\mathbf{X}$ is positive definite and thus invertible. We obtain the unique solution $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}$.

(b) Show that $\widehat{\boldsymbol{\beta}}$ is unbiased and that the variance of $\widehat{\boldsymbol{\beta}}$ is given by $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

Solution: The proof of unbiasedness is trivial. For the variance,

$$\begin{aligned}
\mathrm{var}(\widehat{\boldsymbol{\beta}}) &= \mathrm{var}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y}\} \\
&= \mathrm{var}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\beta + \boldsymbol{\varepsilon})\} \\
&= \mathbf{0} + \mathrm{var}\{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{var}(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.
\end{aligned}$$

Exercise 1.3 Linear regression for binary classification. Consider a binary classification problem with $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Y} = \mathcal{A} = \{-1, 1\}$. We model the conditional expectation of Y given $\boldsymbol{X}$ by the linear model $\mathbb{E}(Y \mid \boldsymbol{X}) = \boldsymbol{X}^\top\beta$.

Let $\boldsymbol{x} \in \mathbb{R}^n$ be a new input. So, we estimate $\widehat{\mathbb{E}}(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is the least-square estimate of β. We wish to estimate its class $y = f^*(\boldsymbol{x})$, where f^* is the target function corresponding to $0-1$ loss.

(a) Derive the linear model estimate of $\widehat{\mathbb{P}}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$.

Solution: Show that $\mathbb{E}(Y \mid \boldsymbol{X}) = 2\mathbb{P}(Y = 1 \mid \boldsymbol{X}) - 1$. Hence $\mathbb{P}(Y = 1 \mid \boldsymbol{X}) = \{\mathbb{E}(Y \mid \boldsymbol{X}) + 1\}/2$ and

$$\widehat{\mathbb{P}}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\boldsymbol{x}^\top\widehat{\boldsymbol{\beta}} + 1}{2}.$$

(b) Show that $\widehat{y} = \widehat{f}^*(\boldsymbol{x})$ is given by $2 \cdot 1\{\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}} \geq 0\} - 1$, where $\widehat{f}^*$ is the estimate of f^* given by plugging-in estimated values $\widehat{\mathbb{P}}(Y = y \mid \boldsymbol{X} = \boldsymbol{x})$ of the conditional p.m.f. $\mathbb{P}(Y = y \mid \boldsymbol{X} = \boldsymbol{x})$.

Solution: We will predict $\widehat{y} = 1$ when $\widehat{\mathbb{P}}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) \geq 1/2$, and $\widehat{y} = -1$ when $\widehat{\mathbb{P}}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1/2$. This gives the result.

Exercise 1.4 Pinball loss and quantile regression. For $\tau \in]0, 1[$, the *pinball function* with parameter τ is the function h_τ given by,

$$h_\tau(z) = -\tau\, z\, 1_{\{z \leq 0\}} \;+\; (1 - \tau)\, z\, 1_{\{z > 0\}}.$$

We consider a decision problem for which inputs, outputs and actions are all real-valued, that is $\mathcal{X} = \mathcal{Y} = \mathcal{A} = \mathbb{R}$. For $a, y \in \mathbb{R}$, we define the pinball loss by $\ell_\tau(a, y) = h_\tau(a - y)$.

We assume further that

(a) $\mathbb{E}[|Y||X = x] < \infty$ a.e. $x \in \mathbb{R}$,

(b) and the conditional law of Y given X is absolutely continuous with respect to the Lebesgue measure. Thus the function $y \mapsto \mathbb{P}(Y \leq y \mid X = x)$ is continuous, a.e. $x \in \mathbb{R}$.

Recall that for a real-valued random variable Y whose law is absolutely continuous, we define the quantile of order α or α-quantile as the unique $q_\alpha \in \mathbb{R}$ such that $\mathbb{P}(Y \leq q_\alpha) = \alpha$. Similarly, the conditional quantile of order α of Y at $X = x$ is, under the above continuity hypothesis the unique $q_\alpha(x) \in \mathbb{R}$ such that

$$\mathbb{P}(Y \leq q_\alpha(x) \mid X = x) = \alpha.$$

(a) Plot the pinball function in R. Play around with different values of τ. Why do you think the function is called that way?

(b) Compute the expression for the conditional risk associated with the pinball loss in terms of q_α.

Solution: Let F_x denote the c.d.f. of Y conditional on $X = x$. Then,

$$\begin{aligned}
\mathcal{R}(a|x) &= \mathbb{E}\left[h_\tau(a - Y) \mid X = x\right] \\
&= \int_{\mathbb{R}} h_\tau(a - y) dF_x(y) \\
&= \int_0^1 h_\tau(a - q_\alpha(x)) d\alpha
\end{aligned}$$

by the substitution $y \mapsto q_\alpha(x)$.

(c) Prove that the target function of that risk is $q_\tau(x)$.

Solution: Notice that for $z \neq 0$, $h'_\tau(z) = 1_{\{z > 0\}} - \tau$. Thus,

$$\begin{aligned}
\tfrac{d}{da}\mathcal{R}(a|x) &= \int_{\mathbb{R}} h'_\tau(a - y) dF_x(y) \\
&= \int_{\mathbb{R}} (1_{\{a > Y\}} - \tau) dF_x(y) \\
&= F_x(a) - \tau
\end{aligned}$$

Clearly, for $a < q_\tau(x)$, the conditional risk is decreasing while for $a > q_\tau(x)$ it is increasing, so it is minimum at $a = q_\tau(x)$. Thus the target function is $q_\tau(x)$.

(d) We call ℓ_1-regression or least absolute deviation regression, the regression with loss function $\ell(a, y) = |a - y|$. Deduce from the previous question what is the target function for ℓ_1-regression.

Solution: Clearly, $|z| = 2h_{1/2}(z)$. It follows that the target function is $q_{1/2}(x)$.

Practical exercises

Exercise 1.5 Polynomial regression. In this exercise, we will fit a linear model to data from `simreg1train.csv`. In R, use the `read.csv("...")` function to import the data.

(a) Using results from Exercise 1.2, compute the least squares estimates for this dataset using your statistical software and plot the fitted values. Is the model appropriate?

(b) Calculate the empirical risk on the training set (also called *training error*) for this dataset, given by

$$\widehat{\mathcal{R}}(\widehat{f}) = \tfrac{1}{n} \sum_{i=1}^{n} \ell(y_i, \widehat{f}(x_i)), \tag{1}$$

where $\{(x_i, y_i)\}_{i=1}^n$ is the training set, ℓ is the squared error loss and $\widehat{f}$ is the fitted function.

(c) For the same loss function, calculate the empirical risk on the testing set (also called *testing error*) which is also given by (1) but here $\{(x_i, y_i)\}_{i=1}^n$ is the testing set given in `simreg1test.csv`.

(d) We now make the model more flexible by adding features to the design matrix $\mathbf{X}$. Add the feature $\boldsymbol{x}^2$ into your regression model, i.e., our design matrix becomes $\mathbf{X} = (\mathbf{1}\ \boldsymbol{x}\ \boldsymbol{x}^2)$. Compute the empirical risks on the training and testing sets for this model. Discuss.

Solution [(a)-(d)]: See the R-code uploaded on moodle.

(e) Add features up to $\boldsymbol{x}^k$ into your regression model, for $k = 3, 4, \dots, 10$. Calculate the the empirical risks on the training and testing sets for each $k = 1, \dots, 10$. Make a plot of the empirical risks against k. Discuss. What happens when $k > 10$?

Solution: See Figure 1. We see that the training error decreases with increasing k. However, the test error decreases initially but increases again after a certain k.

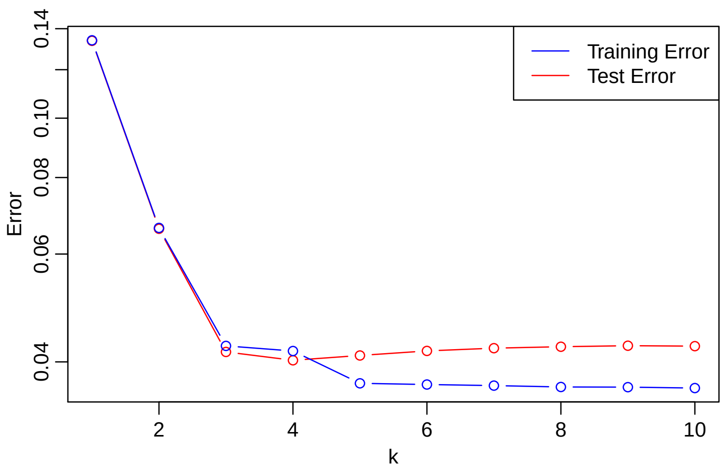

Figure 1: Training and test errors over k. Note the log scale on the y-axis.