EPFL, Autumn 2017
Emeric Thibaud | Jonathan Koh

Statistical Machine Learning

Exercise sheet 4

Exercise 4.1 (Kullblack–Leibler divergence) The Kullback–Leibler (KL) divergence between two densities p and q is defined as

$$\mathrm{KL}(p \mid\mid q) = -\int p(y) \log \frac{q(y)}{p(y)} \mathrm{d}y.$$

Note that KL is not symmetric, $\mathrm{KL}(p \mid\mid q) \neq \mathrm{KL}(q \mid\mid p)$.

(a) Show that $\mathrm{KL}(p \mid\mid q) \geq 0$ with equality if and only if $p(y) = q(y)$ for all $y \in \mathbb{R}$.

Hint: Jensen's inequality says that if the function f is convex then $\mathrm{E}\{f(Y)\} \geq f\{\mathrm{E}(Y)\}$, with strong inequality if f is strictly convex.

(b) Consider a family of density $\{g_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. Show that, for any density p, $\mathrm{KL}(p \mid\mid g_{\boldsymbol{\theta}})$ is minimized for

$$\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} \int p(y) \log g_{\boldsymbol{\theta}}(y) \mathrm{d}y = \arg\max_{\boldsymbol{\theta}} \mathrm{E}_{Y\sim p}\{\log g_{\boldsymbol{\theta}}(Y)\}.$$

(c) Let $\boldsymbol{y} = (y_1, \dots, y_n)$ be a sample of n independent observations from the distribution p. We would like to find the best model $g_{\boldsymbol{\theta}_0}$ for these data by maximizing $K(\boldsymbol{\theta}) = \mathrm{E}_{Y\sim p}\{\log g_{\boldsymbol{\theta}}(Y)\}$, i.e., we want to find $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \mathrm{KL}(p \mid\mid g_{\boldsymbol{\theta}})$. Note that the function $K(\boldsymbol{\theta})$ is unknown because it depends on p.

Write the empirical estimator $\widehat{K}(\theta)$ of $\mathrm{E}_{Y\sim p}\{\log g_{\boldsymbol{\theta}}(Y)\}$ based on $\boldsymbol{y} = (y_1, \dots, y_n)$ and characterize the estimator $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \widehat{K}(\boldsymbol{\theta})$.

(d) Why is $\widehat{K}(\widehat{\boldsymbol{\theta}})$ a bad estimate of $K(\widehat{\boldsymbol{\theta}})$? Akaike proved that (this is not easy!), if $p = g_{\boldsymbol{\theta}_0}$ for some $\boldsymbol{\theta}_0 \in \Theta \subseteq \mathbb{R}^d$ and $d \in \mathbb{N}$,

$$\mathrm{E}_{\boldsymbol{y}}\{\widehat{K}(\widehat{\boldsymbol{\theta}})\} \approx \mathrm{E}_{\boldsymbol{y}}\{K(\widehat{\boldsymbol{\theta}})\} + \frac{d}{n}, \qquad \text{as } n \to \infty.$$

Use Akaike's result to propose an estimate of $K(\widehat{\boldsymbol{\theta}})$. Discuss how this estimator can be used for model selection.

Exercise 4.2 (Practical: variable selection and regularization) You will perform subset selection, ridge and lasso on the bodyfat dataset.

If you work on R, the data are in the library "mfp" and are then loaded using the command `data(bodyfat)`; be careful, you must remove the columns 1,2 and 4 when fitting your models. If you work on another software, you can find the bodyfat data in a csv file on Moodle.

(a) If you haven't already, read §§6.5,6.6 from ISL to familiarize yourself with the packages and the functions in R you will need for the purpose of this exercise.

(b) Perform best-subset selection for $k = 1, \ldots, 13$. Plot the residual sum of squares of the best models over k.

(c) Perform forward and backward subset selection for $k = 1, \ldots, 13$.

(d) Perform ridge and lasso regression for some sequences of λ. By using `plot.glmnet`, plot a graph of the coefficient values over λ.

(e) You will try to find optimal values of k for subset selection, and λ for ridge and lasso, by evaluating the errors of the models on an independent test set. Unfortunately, we don't have additional observations, you will then split your original dataset in two parts: 152 observations used for the training set and 100 observations used for the test set. This split should be done at random, but it is useful to fix the random seed of R first using for example `set.seed(1)`. You can use the command `sample(1:252,152)` to get indices of 152 observations to be used in the training set.

Run best-subset, forward and backward selection procedures and identify for each method the value of k that minimizes the test error.

Similarly, for ridge and lasso, identify a good value of λ that minimizes the test error. You can plot a graph of the test error over a sequence of values of λ.

(f) What happens when you fix a different random seed instead? i.e., repeat the procedure by first calling `set.seed(5)`, this will change the training and test sets. Does your result from (e) change? Discuss.