Exercise Collection – Advanced Risk Min

Contents

Lecture exercises

Exercise 1: Risk Minimization and Gradient Descent (Part 1)

You want to estimate the relationship between a continuous response variable $\boldsymbol{y} \in \mathbb{R}^n$ and some feature $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ using the linear model with an appropriate loss function L.

(a) Describe the model f used in this case, its hypothesis space $\mathcal{H}$ and the theoretical risk function.

(b) Given $f \in \mathcal{H}$, explain the different parts of the Bayes regret if (i) $f^* \in \mathcal{H}$; if (ii) $f^* \notin \mathcal{H}$.

(c) Define the empirical risk and derive the gradients of the empirical risk.

(d) Show that the empirical risk is convex in the model coefficients. Why is convexity a desirable property? Hint: Compute the Hessian matrix $\boldsymbol{H} \in \mathbb{R}^{p \times p}$ and show that $\boldsymbol{z}^\top \boldsymbol{H} \boldsymbol{z} \geq 0 \, \forall \boldsymbol{z} \in \mathbb{R}^p$, i.e., show that the Hessian is positive semi-definite (psd).

(e) Write a function implementing a gradient descent routine for the optimization of this linear model. Start with:

```
#' @param step_size the step_size in each iteration
#' @param X the feature input matrix X
#' @param y the outcome vector y
#' @param beta a starting value for the coefficients
#' @param eps a small constant measuring the changes in each update step.
#' Stop the algorithm if the estimated model parameters do not change
#' more than \code{eps}.

#' @return a set of optimal coefficients beta
gradient_descent <- function(step_size, X, y, beta, eps = 1e-8){

  # >>> do something <<<

  return(beta)
}
```

(f) Run a small simulation study by creating 20 data sets as indicated below and test different step sizes α (fixed across iterations) against each other and against the state-of-the-art routine for linear models (in R, using the function `lm`, in Python, e.g., `sklearn.linear_model.LinearRegression`).

- Compare the difference in estimated coefficients $\beta_j, j = 1, \ldots, p$ using the mean squared error, i.e.

$$p^{-1}\sum_{j=1}^{p}(\beta_j^{truth} - \hat{\beta}_j)^2$$

  and summarize the difference over all 100 simulation repetitions.
- Compare the run times of your implementation and the one given by the state-of-the-art method by wrapping the function calls into a timer (e.g., `system.time()` in R).

```r
# settings
n <- 10000
p <- 100
nr_sims <- 20

# create data (only once)
X <- matrix(rnorm(n*p), ncol=p)
beta_truth <- runif(p, -2, 2)
f_truth <- X%*%beta_truth

# create result object
result_list <- vector("list", nr_sims)

for(sim_nr in nr_sims)
{

  # create response
  y <- f_truth + rnorm(n, sd = 2)

  # >>> do something <<<


  # save results in list (performance, time)
  result_list[[sim_nr]] <- add_something_meaningful_here

}
```

(g) Why is gradient descent maybe not the best option for optimization of a linear model with $L2$ loss? What other options exist? Name at least one and describe how you would apply this for the above problem.

(h) Can we say something about the algorithm's consistency w.r.t. $\mathbb{P}_{xy}$, if $f^* \notin \mathcal{H}$?

Exercise 2: Risk Minimization and Gradient Descent (Part 2)

This exercise builds upon the previous exercise sheet.

(a) Write a function implementing a gradient descent routine for the optimization of the linear model defined in the previous exercise sheet. Start with:

```r
#' @param step_size the step_size in each iteration
#' @param X the feature input matrix X
#' @param y the outcome vector y
#' @param beta a starting value for the coefficients
#' @param eps a small constant measuring the changes in each update step.
#' Stop the algorithm if the estimated model parameters do not change
#' more than \code{eps}.

#' @return a set of optimal coefficients beta
gradient_descent <- function(step_size, X, y, beta, eps = 1e-8){

  # >>> do something <<<

  return(beta)

}
```

(b) Run a small simulation study by creating 20 data sets as indicated below and test different step sizes α (fixed across iterations) against each other and against the state-of-the-art routine for linear models (in R, using the function `lm`, in Python, e.g., `sklearn.linear_model.LinearRegression`).

- Compare the difference in estimated coefficients $\beta_j, j = 1, \dots, p$ using the mean squared error, i.e.

$$p^{-1} \sum_{j=1}^{p} (\beta_j^{truth} - \hat{\beta}_j)^2$$

  and summarize the difference over all 20 simulation repetitions.

- Compare the run times of your implementation and the one given by the state-of-the-art method by wrapping the function calls into a timer (e.g., `system.time()` in R).

```r
# settings
n <- 10000
p <- 100
nr_sims <- 20

# create data (only once)
X <- matrix(rnorm(n*p), ncol=p)
beta_truth <- runif(p, -2, 2)
f_truth <- X%*%beta_truth

# create result object
result_list <- vector("list", nr_sims)

for(sim_nr in nr_sims)
{

  # create response
  y <- f_truth + rnorm(n, sd = 2)

  # >>> do something <<<


  # save results in list (performance, time)
  result_list[[sim_nr]] <- add_something_meaningful_here

}
```

(c) Why is gradient descent maybe not the best option for optimization of a linear model with $L2$ loss? What other options exist? Name at least one and describe how you would apply this for the above problem.

(d) Can we say something about the algorithm's consistency w.r.t. $\mathbb{P}_{xy}$, if $f^* \notin \mathcal{H}$?

Exercise 3: Risk Minimizers for 0-1-Loss

Consider the classification learning setting, i.e., $\mathcal{Y} = \{1, \ldots, g\}$, and the hypothesis space is $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$. The loss function of interest is the 0-1-loss:

$$L\left(y, h(\mathbf{x})\right) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}} = \begin{cases} 1, & \text{if } y \neq h(\mathbf{x}), \\ 0, & \text{if } y = h(\mathbf{x}). \end{cases}$$

(a) Consider the hypothesis space of constant models $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y} \,|\, h(\mathbf{x}) = \boldsymbol{\theta} \in \mathcal{Y} \ \forall \mathbf{x} \in \mathcal{X}\}$. where $\mathcal{X}$ is the feature space. Show that

$$\hat{h}(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\}$$

is the empirical risk minimizer for the 0-1-loss in this case.

(b) What is the optimal constant model in terms of the (theoretical) risk for the 0-1-loss and what is its risk?

(c) Derive the approximation error if the hypothesis space $\mathcal{H}$ consists of the constant models.

(d) Assume now $g = 2$ (binary classification) and consider now the hypothesis space of probabilistic classifiers $\mathcal{H} = \{\pi : \mathcal{X} \to [0, 1]\}$, that is, $\pi(\mathbf{x})$ (or $1 - \pi(\mathbf{x})$) is an estimate of the posterior distribution $p_{y|x}(1 \mid \mathbf{x})$ (or $p_{y|x}(0 \mid \mathbf{x})$). Further, consider the probabilistic 0-1-loss

$$L\left(y, \pi(\mathbf{x})\right) = \begin{cases} 1, & \text{if } (\pi(\mathbf{x}) \geq 1/2 \text{ and } y = 0) \text{ or } (\pi(\mathbf{x}) < 1/2 \text{ and } y = 1), \\ 0, & \text{else.} \end{cases}$$

Is the minimum of $\mathbb{E}_{xy}[L\left(y, \pi(\mathbf{x})\right)]$ unique over $\pi \in \mathcal{H}$[1]? Is the posterior distribution $p_{y|x}$ a resp. *the* minimizer of $\mathbb{E}_{xy}[L\left(y, \pi(\mathbf{x})\right)]$? Discuss the corresponding (dis-)advantages of your findings.

Hint: First note that we can write $L\left(y, \pi(\mathbf{x})\right) = \mathbb{1}_{\{\pi(\mathbf{x}) \geq 1/2\}} \mathbb{1}_{\{y=0\}} + \mathbb{1}_{\{\pi(\mathbf{x}) < 1/2\}} \mathbb{1}_{\{y=1\}}$ and then consider the "unraveling trick": $\mathbb{E}_{xy}[L\left(y, \pi(\mathbf{x})\right)] = \mathbb{E}_x\left[\mathbb{E}_{y|x}[L\left(y, \pi(\mathbf{x})\right) \mid \mathbf{x} = \mathbf{x}]\right]$.

Exercise 4: Risk Minimizers for Generalized L2-Loss

Consider the regression learning setting, i.e., $\mathcal{Y} = \mathbb{R}$, and assume that your loss function of interest is $L\left(y, f(\mathbf{x})\right) = \left(m(y) - m(f(\mathbf{x}))\right)^2$, where $m : \mathbb{R} \to \mathbb{R}$ is a continuous strictly monotone function.

Disclaimer: In the following we always assume that $\mathsf{Var}(m(Y))$ exists.

(a) Consider the hypothesis space of constant models $\mathcal{H} = \{f : \mathcal{X} \to \mathbb{R} \,|\, f(\mathbf{x}) = \boldsymbol{\theta} \ \forall \mathbf{x} \in \mathcal{X}\}$, where $\mathcal{X}$ is the feature space. Show that

$$\hat{f}(\mathbf{x}) = m^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} m(y^{(i)})\right)$$

is the optimal constant model for the loss function above, where m^{-1} is the inverse function of m.

Hint: We can obtain several different notions of a mean value by using a specific function m, e.g., the arithmetic mean by $m(x) = x$, the harmonic mean by $m(x) = 1/x$ (if $x > 0$) or the geometric mean by $m(x) = \log(x)$ (if $x > 0$).

[1] If it is unique, then the loss is a strictly proper scoring rule.

(b) Verify that the risk of the optimal constant model is $\mathcal{R}_L\left(\hat{f}\right) = \left(1 + \frac{1}{n}\right) \mathsf{Var}(m(y))$.

(c) Derive that the risk minimizer (Bayes optimal model) f^* is given by $f^*(\mathbf{x}) = m^{-1}\left(\mathbb{E}_{y|x}\left[m(y) \mid \mathbf{x}\right]\right)$.

(d) What is the optimal constant model in terms of the (theoretical) risk for the loss above and what is its risk?

(e) Recall the decomposition of the Bayes regret into the estimation and the approximation error. Show that the former is $\frac{1}{n}\mathsf{Var}(m(y))$, while the latter is $\mathsf{Var}\Big(\mathbb{E}_{y|x}\left[m(y) \mid \mathbf{x}\right]\Big)$ for the optimal constant model $\hat{f}(\mathbf{x})$ if the hypothesis space $\mathcal{H}$ consists of the constant models.

Hint: Use the law of total variance, which states that $\mathsf{Var}(Y) = \mathbb{E}_X\left[\mathsf{Var}(Y \mid X)\right] + \mathsf{Var}(\mathbb{E}_{Y|X}\left[Y \mid X\right])$, where the conditional variance is defined as $\mathsf{Var}(Y \mid X) = \mathbb{E}_X\left[\left(Y - \mathbb{E}_{Y|X}(Y \mid X)\right)^2 \mid X\right]$.

Exercise 5: Connection between MLE and ERM (Part 1)

Imagine you work at a car dealer and are tasked with predicting the monthly number of cars that will be sold within the next year. You decide to address this challenge in a data-driven manner and develop a model that predicts the number of cars from data regarding vehicles' properties from sold cars of previous years, current competitor and market data.

a) Let x_1 and x_2 measure the number of sold cars of the previous month and of the previous year, respectively. Both features and target are numeric and discrete. You choose to use a generalized linear model (GLM) for this task. For this, you assume the targets to be conditionally independent given the features, i.e., $y^{(i)}|\mathbf{x}^{(i)} \perp y^{(j)}|\mathbf{x}^{(j)}$ for all $i, j \in \{1, 2, \ldots, n\}, i \neq j$, with sample size n.

- Argue which of the following distributions from the one-parametric exponential family is most suitable for the underlying use case: normal, Bernoulli, gamma or Poisson.
- Write down the probability distribution of the chosen distribution depending on $\boldsymbol{\theta}$ assuming a log link function.

b) State the hypothesis space for the corresponding model class. For this, assume the parameter vector $\boldsymbol{\theta}$ to include the intercept coefficient.

c) Which parameters need to be learned? Define the corresponding parameter space Θ.

d) In classical statistics, you would estimate the parameters via maximum likelihood estimation (MLE).

Describe how you can make use of the likelihood in empirical risk minimization (ERM) and write down the likelihood as well as the resulting empirical risk.

Exercise 6: Connection between MLE and ERM (Part 2)

Suppose we are facing a regression task, i.e., $\mathcal{Y} = \mathbb{R}$, and the feature space is $\mathcal{X} \subseteq \mathbb{R}^p$. Let us assume that the relationship between the features and labels is specified by

$$y = m^{-1}\left(m(f_{\text{true}}(\mathbf{x})) + \epsilon\right), \tag{1}$$

where $m : \mathbb{R} \to \mathbb{R}$ is a continuous strictly monotone function with m^{-1} being its inverse function, and the errors are Gaussian, i.e. $\epsilon \sim \mathcal{N}(0, \sigma^2)$. In particular, for the data points $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$ it holds that

$$y^{(i)} = m^{-1}\left(m(f_{\text{true}}(\mathbf{x}^{(i)})) + \epsilon^{(i)}\right), \tag{2}$$

where $\epsilon^{(1)}, \ldots, \epsilon^{(n)}$ are iid with distribution $\mathcal{N}(0, \sigma^2)$.

Disclaimer: We assume in the following that $m(y)$ and $m(f(\mathbf{x}))$ is well-defined for any $y \in \mathcal{Y}$, $f \in \mathcal{H}$ and $\mathbf{x} \in \mathcal{X}$.

(a) How can we transform the labels $y^{(1)}, \dots, y^{(n)}$ to “new” labels $z^{(1)}, \dots, z^{(n)}$ such that $z^{(i)} \mid \mathbf{x}$ is normally distributed? What are the parameters of this normal distribution?

(b) Assume that the hypothesis space is

$$\mathcal{H} = \{f(\cdot \mid \boldsymbol{\theta}) : \mathcal{X} \to \mathbb{R} \ \mid f(\cdot \mid \boldsymbol{\theta}) \text{ belongs to a certain functional family parameterized by } \boldsymbol{\theta} \in \Theta\},$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_d)$ is a parameter vector, which is an element of a **parameter space** Θ. Based on your findings in (a), establish a relationship between minimizing the negative log-likelihood for $(\mathbf{x}^{(1)}, z^{(1)}), \dots, (\mathbf{x}^{(n)}, z^{(n)})$ and empirical loss minimization over $\mathcal{H}$ of the generalized L2-loss function of Exercise sheet 1, i.e., $L\left(y, f(\mathbf{x})\right) = \big(m(y) - m(f(\mathbf{x}))\big)^2$.

(c) In many practical applications such as biology, medicine, physics or social sciences one often observed statistical property is that the label y given a feature $\mathbf{x}$ follows a *log-normal distribution*[2]. Note that we can obtain such a relationship by using $m(x) = \log(x)$ above. In the following we want to consider the conjecture of the Scottish physician James D. Forbes, who conjectured in the year 1857 that the relationship between the air pressure (in inches of mercury) y and the boiling point of water x (in degrees Farenheit) is given by

$$y = \theta_1 \exp(\theta_2 x + \epsilon),$$

for some specific values $\theta_1 \in \mathbb{R}_+, \theta_2 \in \mathbb{R}$ and some error term ϵ (of course, we assume that this error term is stochastic and normally distributed).

- What would be a suitable hypothesis space $\mathcal{H}$ if this conjecture holds?
- The dataset `forbes` in the R-package `MASS` contains 17 different observations of y and x at different locations in the Alps and Scotland, i.e., the data set is $(x^{(i)}, y^{(i)})_{i=1}^{17}$. Analyze whether his conjecture was reasonable by using the following code snippet:

```r
#' @param X the feature input matrix X
#' @param y the outcome vector y
#' @param theta parameter vector for the model (2-dimensional)

# Load MASS and data set forbes
library(MASS)
data(forbes)
attach(forbes)

# initialize the data set
X = cbind(rep(1,17),bp)
y = pres

#' function to represent your models via the parameter vector theta = c(theta_1, theta_2)
#' @return a predicted label y_hat for x
f <- function(x, theta){

  # >>> do something <<<

  return(y_hat)

}

#' @return a vector consisting of the optimal  parameter vector
optim_coeff <- function(X,y){

  # >>> do something <<<

  return(theta)

}
```

[2] The Wikipedia article on the log-normal distribution has quite a large part about the occurrence of the log-normal distribution.

```
# >>>  Do something here to check Forbes' conjecture <<<
```

Hint: As a sanity check whether your function to find the optimal coefficients work, it should hold that $\hat{\theta}_1 \approx 0.3787548$ and $\hat{\theta}_2 \approx 0.02062236$.