Exercises
Introduction to Machine Learning
SS 2019

Series 2, March 4th, 2019 (Regression, Classification)

Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `https://las.inf.ethz.ch/teaching/introml-s19`
Email questions to:
yehuda.levy@inf.ethz.ch, laurie.prelot@inf.ethz.ch

Problem 1 (Regression):

Let $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ be the training data that you are given. To predict y as $\mathbf{w}^T\mathbf{x}$ for some parameter vector $\mathbf{w} \in \mathbb{R}^d$ we can use

The *ordinary least square optimization (OLS)* problem :

$$\operatorname*{argmin}_{\mathbf{w}} \hat{R}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2. \tag{1}$$

The *ridge regression* optimization problem with parameter $\lambda > 0$:

$$\operatorname*{argmin}_{\mathbf{w}} \hat{R}_{\text{ridge}}(\mathbf{w}) = \operatorname*{argmin}_{\mathbf{w}} \left[\sum_{i=1}^{n} \left(y_i - \mathbf{w}^T\mathbf{x}_i\right)^2 + \lambda \mathbf{w}^T\mathbf{w}\right]. \tag{2}$$

We define the OLS and ridge estimator as $\hat{\mathbf{w}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$; $\hat{\mathbf{w}}_{\text{ridge}}(\lambda) = \left(\mathbf{X}^T\mathbf{X} + \lambda I_d\right)^{-1}\mathbf{X}^T\mathbf{y}$ respectively

(a) Show that the ridge penalty shrinks the low variance components, i.e show that it shrinks the singular values.

(b) Show that the ridge regression estimator is biased (*Hint: use the expectation*).
What happens when $\lambda \to \infty$?

(c) Compare the variance of the OLS estimator to that of the ridge regression estimator. How does the variance behave when $\lambda \to \infty$?

Problem 2 (Regression 2):

In this problem you will help Ada solve a linear regression problem. From the domain experts she has learned that it makes sense to use the following regularizer[1],

$$R(\mathbf{w}) = \sum_{i=1}^{d-1} |w_i - w_{i+1}|$$

for the weight vector $\mathbf{w} \in \mathbb{R}^d$. She is given n data points $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$, where each $\mathbf{x}_i \in \mathbb{R}^d$ and each $y_i \in \mathbb{R}$. Hence, she has to *minimize* the following objective

$$f(\mathbf{w}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \underbrace{(\mathbf{w}_i^T\mathbf{x}_i - y_i)^2}_{\text{loss}(\mathbf{w}|y_i,\mathbf{x}_i)}}_{L(\mathbf{w})} + \lambda R(\mathbf{w}).$$

[1] This regularizer makes sense if we would like to prefer solutions whose entries do not change much between adjacent coordinates.

1. Ada wrote a program and then solved the above problem for the *same data points* and four *different* positive penalizers $\lambda_1 < \lambda_2 < \lambda_3 < \lambda_4$. Unfortunately, she has misnamed the files holding the results and does not know which file corresponds to which λ_i. Your task is to help Ada by assigning to each file the corresponding λ_i that was used. Please justify your answer.

File name	Computed weight vector $\mathbf{w}^*$	Penalizer
`solution_a.pkl`	$(1, 1, 2, 2, 1, 1)$	
`solution_b.pkl`	$(9, 10, 10, 8, 2, 2)$	
`solution_c.pkl`	$(2, 2, 4, 5, 5, 5)$	
`solution_d.pkl`	$(1, 2, 2, 2, 3, 1)$	

2. Ada's colleague Alan wrote another program to solve the same optimization problem, but arrived at a different optimum for the same penalizer $\lambda > 0$. Does this mean that one of them has an implementation bug?

3. To ensure that her algorithm is correctly implemented, Ada wants to implement the following test procedure. First, come up with some synthetic distribution $P(\mathbf{x}, y)$ where the data comes from. Then, compute the optimal vector $\mathbf{w}^*$ on a finite sample from $P(\mathbf{x}, y)$, and finally compute the *generalization error* of $\mathbf{w}^*$. If she defined the distribution generating the data as

$$P(\mathbf{x}, y) = \begin{cases} \frac{1}{8} & \text{if } \mathbf{x} \in \{0,1\}^3 \text{ and } y = x_1 + 2x_2 + 2x_3, \text{ or} \\ 0 & \text{otherwise,} \end{cases}$$

and she computed the vector $\mathbf{w}_* = (2, 2, 2)$ on the finite sample, what is the *generalization error*?

Problem 3 (Perceptron):

(a) Construct a perceptron which correctly classifies the following data. Choose appropriate values for the weights $\mathbf{w0}, \mathbf{w1}$ and $\mathbf{w2}$

Training Example	x1	x2	class
a	0	1	-1
b	2	0	-1
c	1	1	+1

(b) Use the perceptron learning algorithm on the data above, using a learning rate ν of 1.0 and initial weight values of
$\mathbf{w0} = -0.5, \mathbf{w1} = 0$ and $\mathbf{w2} = 1$
You can fill this table :

Iteration i	w0	w1	w2	Training Example (a, b or c)	current class	s=w0+w1x1+w2x2	Action