Exercises
**Introduction to Machine Learning**
SS 2019

**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:**
laurie.prelot@inf.ethz.ch, yehuda.levy@inf.ethz.ch

# Series 2, March 4th-8th, 2019 (Regression, Classification)

**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 ridge estimator as $\hat{\mathbf{w}}_{\text{ridge}}(\lambda) = \left(\mathbf{X}^T\mathbf{X} + \lambda I_d\right)^{-1}\mathbf{X}^T\mathbf{y}$

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

*Solution*:
Both the OLS and the ridge estimators can be rewritten in term of the SVD matrices.

$$\begin{aligned}
\hat{\mathbf{w}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\right)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{y} \\
&= \left(\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T\right)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{V}\boldsymbol{\Sigma}^{-2}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{V}\boldsymbol{\Sigma}^{-2}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{y}
\end{aligned}$$

$$\begin{aligned}
\hat{\mathbf{w}}_{\text{ridge}}(\lambda) &= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T + \lambda\mathbf{I}\right)^{-1}\mathbf{V}\Sigma\mathbf{U}^T\mathbf{y} \\
&= \mathbf{V}\left(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I}\right)^{-1}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{y} \\
&= \mathbf{V}\left(\boldsymbol{\Sigma}^2 + \lambda\mathbf{I}\right)^{-1}\boldsymbol{\Sigma}\mathbf{U}^T\mathbf{y}
\end{aligned}$$

Writing $\boldsymbol{\Sigma}_{jj} = d_{jj}$ we have: $d_{jj}^{-1} \geq \frac{d_{jj}}{d_{jj}^2 + \lambda}$ for all $\lambda > 0$

Thus, the ridge penalty will shrink the singular values

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

*Solution*:
Let us study the expectation of the ridge estimator

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathbf{w}}_{\text{ridge}}(\lambda)\right] &= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}\right] \\
&= \mathbb{E}\left[\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\hat{\mathbf{w}}\right] \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\mathbb{E}\left(\hat{\mathbf{w}}\right) \\
&= \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{w}
\end{aligned}$$

We can see that $\mathbb{E}\left[\hat{\mathbf{w}}_{\text{ridge}}(\lambda)\right] \neq \mathbf{w}$ for any $\lambda > 0$ . Hence, the ridge estimator is biased

Then, when $\lambda \to \infty$ :

$$\lim_{\lambda\to\infty} \mathbb{E}\left[\hat{\mathbf{w}}_{\text{ridge}}(\lambda)\right] = \lim_{\lambda\to\infty} \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{w} = 0_d$$

All the regression coefficients are shrunken towards zero as the penalty parameter increases.

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

*Solution*:
As calculated above, we have: $\hat{\mathbf{w}}_{\text{ridge}}(\lambda) = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\hat{\mathbf{w}}$
We define: $\Omega_\lambda = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)$
It can be seen that,

$$\begin{aligned}
Var\left[\hat{\mathbf{w}}_{\text{ridge}}(\lambda)\right] &= Var\left[\Omega_\lambda\hat{\mathbf{w}}\right] \\
&= \Omega_\lambda Var\left[\hat{\mathbf{w}}\right]\Omega_\lambda^T \\
&= \sigma^2\Omega_\lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\Omega_\lambda^T \\
&= \sigma^2\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\left[\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\right]^T
\end{aligned}$$

Note that we have used the fact that $Var\left(\mathbf{AY}\right) = \mathbf{A}Var(\mathbf{Y})\mathbf{A}^T$ for a non random matrix $\mathbf{A}$ , and the fact that $Var\left(\hat{\mathbf{w}}\right) = \sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$

We can now compare it to the variance of the OLS estimator

$$\begin{aligned}
Var\left[\hat{\mathbf{w}}\right] - Var\left[\hat{\mathbf{w}}_{\text{ridge}}(\lambda)\right] &= \sigma^2 \left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1} - \Omega_\lambda \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\Omega_\lambda^T\right] \\
&= \sigma^2 \Omega_\lambda \left[\left(\mathbf{I} + \lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{I} + \lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^T - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]\Omega_\lambda^T \\
&= \sigma^2 \Omega_\lambda \left[2\lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-2} + \lambda^2\left(\mathbf{X}^T\mathbf{X}\right)^{-3}\right]\Omega_\lambda^T \\
&= \sigma^2 \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\left[2\lambda\mathbf{I} + \lambda^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right]\left[\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\right]^T
\end{aligned}$$

The difference is non-negative definite. Hence, the variance of the OLS estimator exceeds that of the ridge estimator.

$$Var\left[\hat{\mathbf{w}}\right] \succeq Var\left[\hat{\mathbf{w}}_{\text{ridge}}(\lambda)\right]$$

Now, let us look at the case where $\lambda \to \infty$ :

$$\lim_{\lambda\to\infty} Var\left[\hat{\mathbf{w}}_{\text{ridge}}(\lambda)\right] = \lim_{\lambda\to\infty} \sigma^2\Omega_\lambda\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\Omega_\lambda^T = 0_d$$

The variance of the ridge estimator vanishes. Hence, the variance of the ridge regression coefficient estimates decreases towards zero as the penalty parameter becomes large.

**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. 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 $\lambda_i$. Your task is to help Ada by assigning to each file the corresponding $\lambda_i$ 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 | $(1, 2, 2, 2, 3, 1)$ | |

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

*Solution*: Take any $\mathbf{w}$ and $\mathbf{w}'$ satisfying $R(\mathbf{w}) < R(\mathbf{w}')$ that are optimal for some $\lambda \neq \lambda'$. Then, because they are optimal for the corresponding losses

$$L(\mathbf{w}) + \lambda R(\mathbf{w}) \leq L(\mathbf{w}') + \lambda R(\mathbf{w}'), \text{ and}$$
$$-L(\mathbf{w}) - \lambda' R(\mathbf{w}) \leq -L(\mathbf{w}') - \lambda' R(\mathbf{w}').$$

Adding both equations we have $(\lambda - \lambda')R(\mathbf{w}) \leq (\lambda - \lambda')R(\mathbf{w}')$. Because $R(\mathbf{w}) \leq R(\mathbf{w}')$, the above is satisfied if $\lambda \geq \lambda'$, and this inequality has to be strict as $\lambda \neq \lambda'$ by assumption.

Because the regularizer for the four parameter vectors evaluates to 2, 9, 3 and 4 respectively, this means that the order is $\lambda_4, \lambda_1, \lambda_3, \lambda_2$.

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?

   *Solution*: No it does not, consider the case where all $\mathbf{x}_i$ and all $y_i$ are equal to zero. Then any constant vector is a solution.

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*?

   *Solution*: Note that there will be no loss if $x_1 = 0$, since in this case $\mathbf{w}_*{}^\top x = y$. On the other hand if $x_1 = 1$ then the loss is always $1$ irrespective of the values of $x_2$ and $x_3$, since in this case $\mathbf{w}_*{}^\top x = 2x_1 + 2x_2 + 2x_3 = x_1 + y = 1 + y$. Hence, the expected loss is equal to $1 \cdot P(x_1 = 1) = \frac{1}{2}$.

**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 |

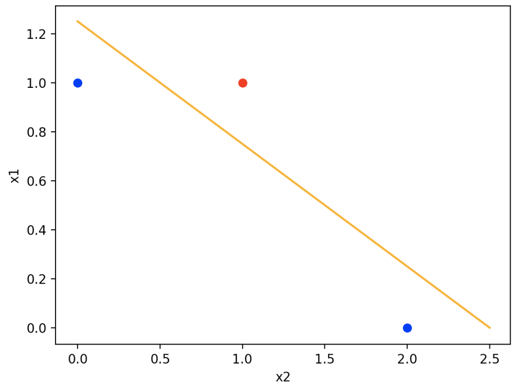
*Solution*: We can plot the data and trace a separation line. This line has slope -1/2 and x2-intersect 5/4. $x2 = 5/4 - x1/2$ i.e. $2x1 + 4x2 - 5 = 0$ Thus we can choose , $w0 = -5, w1 = 2, w2 = 4$

(b) Use the perceptron learning algorithm on the data above, using a learning rate $\nu$ 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 ) | Class | s=w0+w1x1+w2x2 | Action |
|---|---|---|---|---|---|---|---|
| | | | | | | | |

*Solution*: We apply stochastic gradient descent. To facilitate this pen and paper exercise, we do not pick a sample at random but will take a, b and c sequentially.

Figure 1: Problem 3 (b), Classification




| Iteration i | w0 | w1 | w2 | Training Example (a, b or c ) | Class | s=w0+w1x1+w2x2 | Action |
|---|---|---|---|---|---|---|---|
| 1 | -0.5 | 0 | 1 | a. | - | 0.5 | Update |
| 2 | -1.5 | 0 | 0 | b. | - | -1.5 | None |
| 3 | -1.5 | 0 | 0 | c. | + | -1.5 | Update |
| 4 | -0.5 | 1 | 1 | a. | - | 0.5 | Update |
| 5 | -1.5 | 1 | 0 | b. | - | 0.5 | Update |
| 6 | -2.5 | -1 | 0 | c. | + | -3.5 | Update |
| 7 | -1.5 | 0 | 1 | a. | - | -0.5 | None |
| 8 | -1.5 | 0 | 1 | b. | - | -1.5 | None |
| 9 | -1.5 | 0 | 1 | c. | + | -0.5 | Update |
| 10 | -0.5 | 1 | 2 | a. | - | 1.5 | Update |
| 11 | -1.5 | 1 | 1 | b. | - | 0.5 | Update |
| 12 | -2.5 | -1 | 1 | c. | + | -2.5 | Update |
| 13 | -1.5 | 0 | 2 | a. | - | 0.5 | Update |
| 14 | -2.5 | 0 | 1 | b. | - | -2.5 | None |
| 15 | -2.5 | 0 | 1 | c. | + | -1.5 | Update |
| 16 | -1.5 | 1 | 2 | a. | - | 0.5 | Update |
| 17 | -2.5 | 1 | 1 | b. | - | -0.5 | None |
| 18 | -2.5 | 1 | 1 | c. | + | -0.5 | Update |
| 19 | -1.5 | 2 | 2 | a. | - | 0.5 | Update |
| 20 | -2.5 | 2 | 1 | b. | - | 1.5 | Update |
| 21 | -3.5 | 0 | 1 | c. | + | -2.5 | Update |
| 22 | -2.5 | 1 | 2 | a. | - | -0.5 | None |
| 23 | -2.5 | 1 | 2 | b. | - | -0.5 | None |
| 24 | -2.5 | 1 | 2 | c. | + | 0.5 | None |