Exercises
Machine Learning
HS 2016

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/ml/

Series 1, Oct 6th, 2016 (Probability and Regression)

Email questions 1, 2 to: **Karim Labib**
labibk@student.ethz.ch
Email questions 3, 4, 5, 6 to: **Qin Wang**
qwang@student.ethz.ch

Please hand in solutions until Friday Oct 21st.

("*"-exercies are a little bit more difficult, but still useful and doable)

Problem 1 (Various Problems):

1. A coin is tossed independently and repeatedly with the probability of heads p.
a) What is the probability of only heads in the first n tosses?
b) What is the probability of obtaining the first tail at the n-th toss?

2. Prove that X independent of Y implies $\mathrm{Cov}(X, Y) = 0$.

3. Show that the variance of a random variable X can be expressed as follows:

$$\mathrm{Var}X = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

4. Let X and Y be such random variables, that $\mathbb{E}X = 2$, $\mathbb{E}Y = 4$, and the following constraint holds true $X^2 + Y = 8$. Find:
a) $\mathbb{E}[X + Y]$
b) $\mathrm{Var}X$

5. Find mean and variance of a continuous uniform $[a, b]$-distribution:

$$p_{\mathsf{unif}}(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

6*. (Weak Law of Large Numbers) Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables, $\mu = \mathbb{E}X_1 < \infty$, $\mathrm{Var}X_1 < \infty$. Prove that the *empirical mean* converges in probability to the true mean:

$$\bar{X}(n) \xrightarrow{\mathbb{P}} \mu \;\; (n \to \infty) \quad \text{where } \bar{X}(n) := \frac{1}{n} \sum X_i.$$

Hint: use the definition of convergence in probability (from the tutorial slides) and the *Chebychev's inequality*:

$$\mathbb{P}\big(|Z - \mathbb{E}Z| \geq \varepsilon\big) \leq \frac{\mathrm{Var}Z}{\varepsilon^2}.$$

Offtopic Note: Chebychev's inequality justifies the usage of variance as a measure of "average deviation" of a random variable from its mean.

Problem 2 (Conditional Probability):

A couple has two children, each of them being independently a boy or a girl with 50% probability. Compute the probabilities of the following events.

1. At least one of the children is a girl.
2. Both children are girls.
3. Both children are girls given that the first born is a girl.
4. Both children are girls given that one of them is a girl.
5. Both children are girls given that one of them is a girl named Cassiopeia.

Note: Cassiopeia is an extremely rare name with a frequency of less than 1 in 1,000,000.

Problem 3 (Regression):

Consider the linear regression model expressed as: $y = \beta_0 + \sum_{i=1}^{D} \mathbf{x}_i \beta_i = \mathbf{x}^\top \beta$, where $x = (1, x_1, ..., x_D) \in \mathbb{R}^{D+1}$ is the input variable and y is the corresponding target variable. Assume that the input dataset is given by the matrix $\mathbf{X} \in \mathbb{R}^{N \times (D+1)}$whose first column is 1. Then the linear regression model for all the observations is written as $y = \mathbf{X}\beta$. Consider this linear regression model and answer the following questions.

1. For this problem, formally define the Residual Sum of Squares(RSS) cost function and write it down in matrix notation.
2. Formally define ridge and LASSO regression models and their cost functions
3. Derive the optimal weight vector $\hat{\beta}_{RSS}$ and $\hat{\beta}_{ridge}$ that minimizes the RSS/Ridge cost function.
4. Computing $\hat{\beta}_{RSS}$ involves the inversion of the symmetric matrix $\mathbf{S} := \mathbf{X}^\top \mathbf{X}$, which we assume to be positive definite. Give a mathematical condition (in terms of the eigenvalues of $\mathbf{S}$), when this inversion is numerically unstable.
5. Explain why choosing $\lambda > 0$ in Ridge regression improves stability of the inversion. What is the effect of increasing λ?

   Weighted Linear Regression

   In standard linear regression all points have equal importance. Now we consider another setup where we pay more attention to specific examples. Specifically, suppose we want to minimize

   $$L_{weighted}(\beta) = \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2$$

   Note that in question 1 we covered the case where $w_i = 1$
6. Write $L_{weighted}(\beta)$ in matrix notation. (Hint: Use a diagonal matrix $\mathbf{W}$ for the weights)
7. Find the optimal β as a function of $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{y}$.

Problem 4 (Bias Variance Decomposition):

Recall that mean squared error(MSE) is defined as $\mathbb{E}_{XY}[(Y - f(X))^2]$

1. Write down the mathematical definition of bias and variance.
2. Expand the mean squared error and write it in the form of variance + squared bias.

Problem 5 (Combination of Individual Regression Models):

We now look at the problem of regression and how the combination of individual regression models can give better results. Left-column figures show the individual regression models, right-column figures show the true target function (solid) and the output of averaging the individual models to obtain a composite model (dashed).

1. The individual regression models (in Figure 1 and Figure 3) have been regularized using a regularization parameter λ, i.e. the cost function had the following form

$$RSS_{Ridge}(\beta) = \sum_{i=1}^{n} (y_i - \phi(x_i)^\top \beta)^2 + \lambda \beta^\top \beta$$

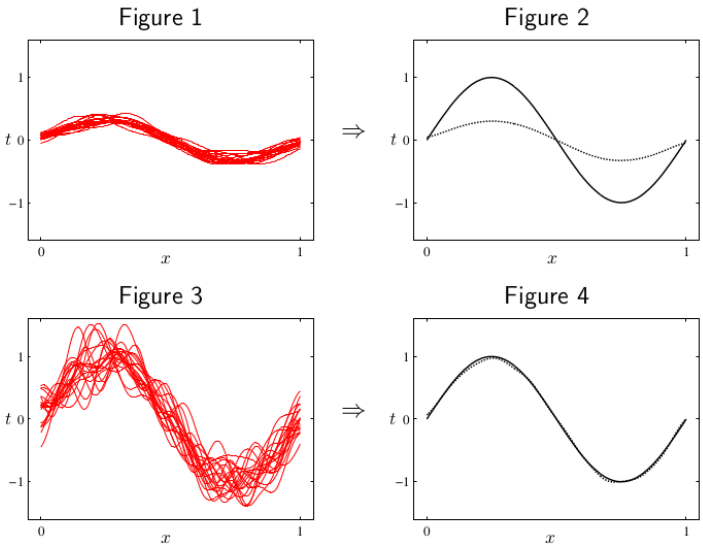
   The parameters used were $\lambda = 0.09$ and $\lambda = 13.5$

   Associate the regularization parameters to the figures in the left column. What is the λ value of Figure 1 and Figure 3.

2. Please interpret the figures in terms of the bias-variance trade-off.

   i. Figure 1 and 2

   ii. Figure 3 and 4



Problem 6 (Python Exercise):

The elastic net is a regularized regression method that linearly combines the L1 and L2 penalties of the LASSO and ridge methods. The cost function of the elastic net method is defined by

$$L_e(w) = \|y - Xw\|_2^2 + \lambda \|w\|_1 + (1 - \lambda)\|w\|_2^2$$

In this exercise, we ask you to perform elastic net analysis on a real-world dataset.

http://archive.ics.uci.edu/ml/datasets/Automobile

Your task is to predict car prices based on 25 given variables. (Read data set description for details)

1. Randomly choose 100 instances as training set, 50 instances as validation set, the rest as test set. You will also need to do some data cleansing work. (Normalization, dealing with missing attribute values...)

2. Use elastic net model to predict prices. How do you determine lambda value? Print out the average loss on the testing set.

3. Can you get better results on the test set by constructing new features?