MAD 2020-21, Assignment 1

Bulat Ibragimov, Kim Steenstrup Pedersen

hand in until: 24.11.2020 at 23:59

General comments: The assignments in MAD must be completed and written individually. You are allowed (and encouraged) to discuss the exercises in small groups. If you do so, you are required to list your group partners in the submission. The report must be written completely by yourself. In order to pass the assignment, you will need to get at least 40% of the available points. The data needed for the assignment can be found in the assignment folder that you download from Absalon.

Submission instructions: Submit your report as a PDF, not zipped up with the rest. Please add your source code to the submission, both as executable files and as part of your report. To include it in your report, you can use the `lstlisting` environment in LaTeX, or you can include a "print to pdf" output in your pdf report.

Exercise 1 (**Partial Derivatives, 3 points).** Compute partial derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ of the following functions:

a) $f(x,y) = x^4y^3 + x^5 - e^y$

b) $f(x,y) = \frac{1}{\sqrt{x^3+xy+y^2}}$

c) $f(x,y) = \frac{x^3+y^2}{x+y}$

Deliverables: For each sub-exercise, provide the derivation steps and result.
Solution:

a) $\frac{\partial f}{\partial x} = 4x^3y^3 + 5x^4$ and $\frac{\partial f}{\partial y} = 3x^4y^2 - e^y$

b) $\frac{\partial f}{\partial x} = -\frac{1}{2}\left(x^3+xy+y^2\right)^{-\frac{3}{2}}\left(3x^2+y\right)$ and $\frac{\partial f}{\partial y} = -\frac{1}{2}\left(x^3+xy+y^2\right)^{-\frac{3}{2}}(x+2y)$

c) $\frac{\partial f}{\partial x} = \frac{(x+y)3x^2-(x^3+y^2)}{(x+y)^2} = \frac{2x^3+3x^2y-y^2}{(x+y)^2}$ and $\frac{\partial f}{\partial y} = \frac{(x+y)2y-(x^3+y^2)}{(x+y)^2} = \frac{y^2+2xy-x^3}{(x+y)^2}$

Instructions:
1 point for each sub-exercise.

Exercise 2 (**Gradients, 3 points).** Let $\bar{x}$ be a vector, A a matrix, $\bar{b}$ a vector, and c a scalar. Compute the gradient ∇f with respect to $\bar{x}$ of the following functions:

a) $f(\bar{x}) = \bar{x}^T\bar{x} + c$

b) $f(\bar{x}) = \bar{x}^T\bar{b}$

c) $f(\bar{x}) = \bar{x}^TA\bar{x} + \bar{b}^T\bar{x} + c$

Deliverables: For each sub-exercise, provide the derivation steps and result.
Solution:

a) Let $\bar{x} = [x_1, \ldots, x_D]$. Since $\bar{x}^T\bar{x} = x_1^2 + x_2^2 + \ldots + x_D^2$, we have $\frac{\partial f}{\partial x_i} = 2x_i$ (the constant c vanishes). Hence, $\nabla f(\bar{x}) = 2\bar{x}$.

b) Let $\bar{b} = [b_1, \ldots, b_D]$ and $\bar{x}$ as above. Since $\bar{x}^T\bar{b} = x_1b_1 + x_2b_2 + \ldots + x_Db_D$, we have $\frac{\partial f}{\partial x_i} = b_i$, so $\nabla f(\bar{x}) = \bar{b}$.

c) $\nabla\left(\bar{x}^TA\bar{x} + \bar{b}^T\bar{x} + c\right) = \nabla\left(x^TA\bar{x}\right) + \nabla\left(\bar{b}^T\bar{x}\right) + \nabla c = \left(A + A^T\right)\bar{x} + \bar{b}$. For the last equality, consider that: the derivative of a sum is the sum of derivatives; the derivative of a constant is zero; $\bar{b}^T\bar{x} = \bar{x}^T\bar{b}$ and we computed the derivative of this term in point b). The only term whose derivative we need to compute is: $\bar{x}^TA\bar{x}$. This can be done by writing the matrix-vector product explicitly.

Instructions:
1 point for each sub-exercise.

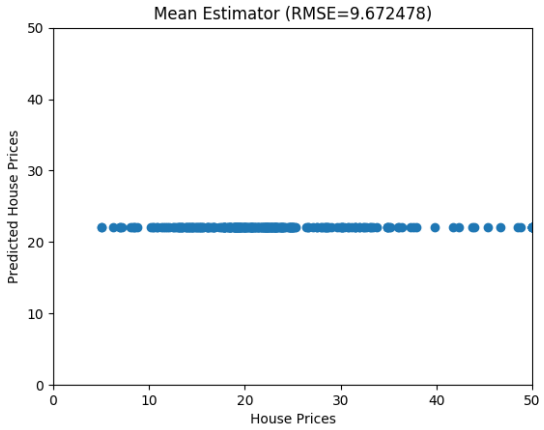

Figure 1: Scatter Plot (Mean Estimator)

Exercise 3 (**Estimating House Prices I, 4 points).** The Boston Housing dataset contains prices for various houses in the area of Boston (Massachusetts, USA) around the year 1978. Your task is to build models that can be used to estimate these prices given information such as the crime rate, the average number of rooms per apartment, average distances to employment centers, and other attributes. You can find a description of the dataset in Appendix A.

The `boston.zip` file that is available on Absalon contains `boston_train.csv` and `boston_test.csv`, which, in turn, contain the training and the test instances, respectively. You are supposed to make use of the training instances to "build" a model; the test instances are used afterwards to evaluate the "quality" of the model.[1]

a) A naive approach to obtain a price estimate for a new house is to resort to the average price of all the houses given in the training set. Implement this simple model, i.e., compute the mean of the house prices given in the **training set**. Extend the `housing_1.py` file—also available on Absalon—that already contains some code to parse both the training and the test dataset.
Hint: Make use of the routines and functions for arrays provided by the Numpy package.

b) Next, evaluate the quality of this model by implementing and using a function `rmse(t,tp)` that computes the root-mean-square error

$$RMSE(\mathbf{t}, \mathbf{t}') = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|t_i - t'_i\|^2}$$

for two N-dimensional vectors $\mathbf{t}$ and $\mathbf{t}'$ (here, `t` and `tp` are arrays containing the values t_i and t'_i). Use this function to compute the RMSE between the true house prices and the estimates obtained via the simple 'mean' model from a) for the instances in the **test set**.

c) Visualize the results by generating a 2D scatter plot ("true house prices" vs. "estimates") for all instances in the **test set** (e.g., via the `matplotlib.pyplot.scatter` function of the `matplotlib` package).

Deliverables: The adapted `housing_1.py` file containing your source code and a) the mean, b) the RMSE, and c) the 2D scatter plot.

Solution:
Code: See reference solution.

a) Mean: 22.01
b) RMSE: 9.67
c) See Figure 1.

Instructions:
2 points for (a) and 1 point for both (b) and (c).

Exercise 4 (**Estimating House Prices II, 4 points).** Next, you will make use of the multivariate linear regression implementation that was discussed during the lecture. More specifically, you are going to fit a linear regression model of the form

$$t = f(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D = \mathbf{w}^T \mathbf{x},$$

[1] Note that the test instances must not be used during the training phase. This mimics a realistic scenario where you do not have access to the target values you aim to predict.

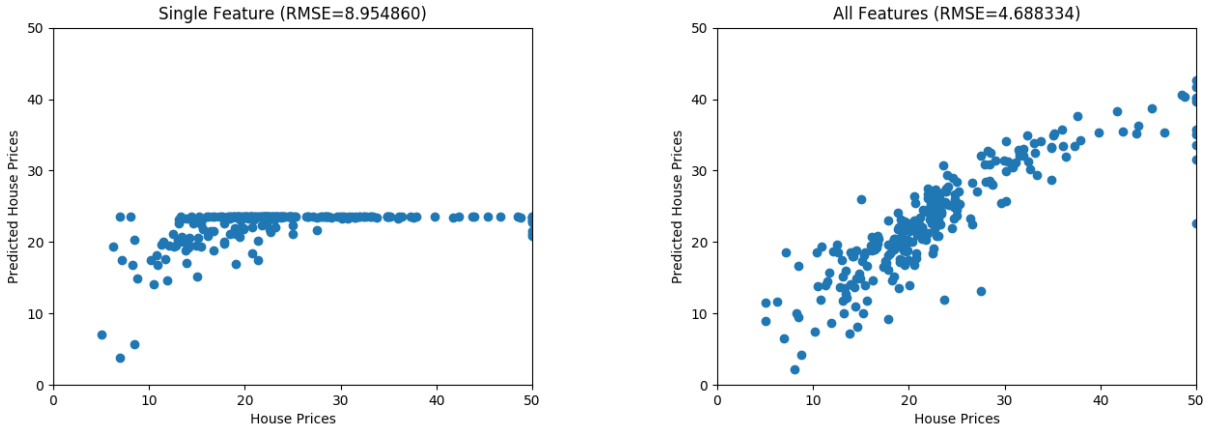

Figure 2: Linear Regression Scatter Plots

where t is the predicted output variable (predicted house price), $\mathbf{x} = (1, x_1, x_2, \ldots, x_D)^T$ is the (augmented) vector-valued input variable, and $\mathbf{w} = (w_0, w_1, \ldots, w_D)$ are the free parameters. The parameters w_i *define* the regression model, and once they have been estimated, the model can be used to predict outputs for a new input vector $\mathbf{x}'$. For the following tasks, make use of the content discussed during the lecture to obtain estimates for these coefficients (i.e., linear modelling via least-squares).

a) We have discussed and developed how to implement linear regression with Python/Numpy. Make use of the corresponding code and (re-)implement linear regression using the supplied template files `housing_2.py` and `linreg.py`: The former one contains some template code and instructions related to parsing the data and addressing the subtasks detailed below. The latter file contains a Python class `LinearRegression` with—so far—incomplete `fit` and `predict` functions. Start by implementing the `fit` function, which shall compute optimal coefficients $\hat{\mathbf{w}}$ (remember the offset parameter w_0). Store the computed weights in the local variable `self.w`; see the template code as well as the code discussed during the lecture for details.

   Note: You should not use a built-in function for computing the optimal weights/fitting the linear regression model. The goal is to make use of standard vector/matrix operations provided by the Numpy package to compute the optimal weights.

b) Use the code developed in a) to fit a linear regression model on a version of the **training set** that only contains *the first feature* (CRIM). What are the two weights? What can you learn from them?

c) Next, fit a model on *all the features* in the **training set**. What are the computed weights?

d) Finally, implement the `predict` function in `linreg.py`, which shall compute, for an array of input instances, the associated predictions. Use the `predict` function to compute predictions for the **test instances** using (i) the model that is based on the first feature only and (ii) the model that is based on all features? What are the two induced RMSE values? Also, generate a 2D scatter plot for each of the two cases ("true house prices" vs. "estimates") based on the test instances.

Deliverables: Both adapted files `housing_2.py` and `linreg.py` containing your source code and b) weights $\hat{w}_0, \hat{w}_1$ and a two-liner, c) weights $\hat{w}_i$, and d) the RMSE for both cases along with two scatter plots.

Solution:
Code: See reference solution. The implementation for linear regression was already discussed during the lecture, which is intended; the students are supposed to adapt the code provided.

b) Optimal weights in case only the first feature is used: $\hat{\mathbf{w}} = (23.63, -0.43)^T$. For each change of 1 unit in x_1, the *predicted* house price t changes -0.43 units (the higher the crime rate, the lower the house price).

c) Optimal weights in case all features are used:

$$\hat{\mathbf{w}} = (31.39, -0.06, 0.03, -0.03, 2.29, -17.33, 3.99, 0.00, -1.29, 0.35, -0.02, -0.81, 0.01, -0.46)$$

d) RMSE using the first feature: 8.95; RMSE using all features: 4.68. The more complex model incorporates more features, which leads to a smaller RMSE. Note that using more features does not necessarily lead to a smaller RMSE (e.g., in case the additional features only contain noise). Scatter plots: See Figure 2.

Instructions:
1 point for each sub-exercise.

Exercise 5 (**Total Training Loss, 2 points, Exercise 1.10 in Rogers & Girolami).** Derive the optimal least squares parameter values $\hat{\mathbf{w}}$ for the total training loss

$$\mathcal{L} = \sum_{n=1}^{N} \left(f(\mathbf{x}_n; \mathbf{w}) - t_n\right)^2 = \sum_{n=1}^{N} \left(\mathbf{w}^T \mathbf{x}_n - t_n\right)^2$$

How does the expression compare with that derived from the average loss? Note: You do not have to prove that the solution is the optimal one (simply providing/computing it is sufficient).

Deliverables: Optimal solution along with its derivation and a short answer to the question.

Solution:
The gradient is given by $\nabla\mathcal{L}(\mathbf{w}) = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{t}$; the solution is therefore given by $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$. Thus, one obtains the same solution as for the average loss. For completeness, one could also mention that the Hessian, given by $\mathbf{H} = \nabla^2 f(\mathbf{w}) = 2\mathbf{X}^T\mathbf{X}$, is positive semidefinite, which means that $\mathcal{L}$ is a convex function and, hence, $\hat{\mathbf{w}}$ the global optimum. Note that convexity was only sketched during the lecture ...
Instructions:
1.5 points if the correct $\hat{\mathbf{w}}$ is derived. 0.5 point for observation that one obtains the same solution. No need to prove that this is a globally optimal solution (i.e., no deduction of points if this is not shown).

Appendix

A The Boston Housing Dataset

A detailed description of the dataset can be found here: `https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`. For the sake of illustration, the main characteristics are given below. On Absalon, you can find a zip file `boston.zip` that contains the following two files:

- `boston_train.csv`: Training set containing 253 instances. Each row contains 14 values separated by commas. The first 13 values correspond to the features; the last value (MEDV) is the target (house price).
- `boston_test.csv`: Corresponding test set containing different 253 instances.

Each row contains the following pieces of information:

- CRIM - per capita crime rate by town
- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $10,000
- PTRATIO - pupil-teacher ratio by town
- B - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in $1000's