DS-GA.1013 Mathematical Tools for Data Science
Homework 3
Yves Greatti - yg390

1. (PCA and linear regression) Consider a dataset of $n$ 2-dimensional data points $x_1, \ldots, x_n \in \mathbb{R}^2$. Assume that the dataset is centered. Our goal is to find a line in the 2D space that lies *closest* to the data. First, we apply PCA and consider the line in the direction of the first principal direction. Second, we fit a linear regression model where $x_i[1]$ is a feature, and $x_i[2]$ the corresponding response. Are these lines the same? Describe each line in terms of the quantity it minimizes geometrically (e.g. sum of some distance from the points to the lines).

   These lines are not the same. The linear regression model approximates the response $x_i[2]$ by finding the feature point $x_i[1]$ the closest to the response point $x_i[2]$ by projecting the response point $x_i[2]$ onto the hyperspace $x[1]^T\beta$. For the OLS estimator this projection is orthogonal: $\beta_{\text{OLS}} = \arg\min_{\beta\in\mathbb{R}} \|x_i[2] - \beta x_i[1]\|_2$.

   PCA finds the principal directions by maximizing the total variance of the sample covariance matrix. Let $u_1$ the first principal direction, $u_1 = \arg\max_{\|v\|_2} \text{Var}(\mathcal{P}_v\mathcal{X})$, where $\mathcal{X}$ is the matrix of $n$ 2-dimensional data points $x_1, \ldots, x_n$. Consider one datapoint $x_i, i = 1, \ldots, n$, applying the Pythagorean theorem we can show that the norm of the total variance squared is: $\|x_i\|^2 = \|\mathcal{P}_{u_1}x_i\|^2 + \|x_i - \mathcal{P}_{u_1}x_i\|^2$ . For a given sample covariance matrix $\mathcal{X}\mathcal{X}^T$, PCA tries to maximize the total variance or equivalently mimize the loss variance by projecting orthogonally the data $\mathcal{X}$ onto the line in the direction of the first principal direction.

2. (Heartbeat) We are interested in computing the best linear estimate of the heartbeat of a fetus in the presence of strong interference in the form of the heartbeat of the baby's mother. To simplify matters, let us assume that we only want to estimate the heartbeat at a certain moment. We have available a measurement from a microphone situated near the mother's belly and another from a microphone that is away from her belly. We model the measurements as

$$\tilde{x}[1] = \tilde{b} + \tilde{m} + \tilde{z_1} \tag{1}$$
$$\tilde{x}[2] = \tilde{m} + \tilde{z_2}, \tag{2}$$

   where $\tilde{b}$ is a random variable modeling the heartbeat of the baby, $\tilde{m}$ is a random variable modeling the heartbeat of the mother, and $\tilde{z_1}$ and $\tilde{z_2}$ model additive noise. From past data, we determine that $\tilde{b}$, $\tilde{m}$, $\tilde{z_1}$, and $\tilde{z_2}$ are all zero mean and uncorrelated with each other. The variances of $\tilde{b}$, $\tilde{z_1}$ and $\tilde{z_2}$ are equal to 1, whereas the variance of $\tilde{m}$ is much larger, it is equal to 10.

   (a) Compute the best linear estimator of $\tilde{b}$ given $\tilde{x}[1]$ in terms of MSE, and the corresponding MSE. Describe in words what the estimator does.

   We have shown in class that centering the variables does not change the MSE, so we want

to estimate MSE = $\min_\beta \mathrm{E}[(\tilde{b} - \beta\tilde{x}[1])^2] = \beta^2\mathrm{Var}(\tilde{x}[1]) + \mathrm{Var}(\tilde{b}) - 2\beta\mathrm{Cov}(\tilde{x}[1], \tilde{b})$.

$$\begin{aligned}
\mathrm{Var}(\tilde{x}[1]) &= \mathrm{Var}(\tilde{b} + \tilde{m} + \tilde{z_1}) \\
&= \mathrm{Var}(\tilde{b}) + \mathrm{Var}(\tilde{m}) + \mathrm{Var}(\tilde{z_1}) \\
&= 1 + 10 + 1 = 12 \\
\mathrm{Cov}(\tilde{x}[1], \tilde{b}) &= \mathrm{E}[\tilde{x}[1]\tilde{b}] \\
&= \mathrm{E}[(\tilde{b} + \tilde{m} + \tilde{z_1})\tilde{b}] = \mathrm{E}[\tilde{b}^2] = 1
\end{aligned}$$

Since $\tilde{b}$, $\tilde{m}$, $\tilde{z}_1$ are all zero mean and uncorrelated with each other.

MSE = $12\beta^2 - 2\beta + 1$, it is a convex quadratic function with respect to $\beta$, so we can set the derivative to zero to find the minimum: $\beta^* = \frac{1}{12}$. $\mathrm{MSE}_{\beta^*} = \frac{11}{12} = 0.91$. This estimator predicts the heartbeat of the baby using the measurement $\tilde{x}[1]$ from the microphone situated near the mother's belly.

(b) Compute the best linear estimator of $\tilde{b}$ given $\tilde{x}$ in terms of MSE, and the corresponding MSE. Describe in words what the estimator does. MSE of this estimator equals $\mathrm{Var}(\tilde{b}) - \Sigma_{\tilde{b}\tilde{x}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{b}\tilde{x}}$.

$$\mathrm{Cov}(\tilde{x}[1], \tilde{x}[2]) = \mathrm{E}[\tilde{x}[1]\tilde{x}[2]] = \mathrm{E}[\tilde{x}[1]]\mathrm{E}[\tilde{x}[2]] = \mathrm{E}[\tilde{x}[1]\tilde{x}[2]]$$

Since $\mathrm{E}[\tilde{x}[1]] = \mathrm{E}[\tilde{x}[2]] = 0$ by linearity of the expectation. And by assumptions

$$\begin{aligned}
\mathrm{Cov}(\tilde{x}[1], \tilde{x}[2]) &= \mathrm{E}[\tilde{x}[1]]\mathrm{E}[\tilde{x}[2]] = \mathrm{E}[\tilde{m}^2] = \mathrm{Var}(\tilde{m}) = 10 \\
\mathrm{Var}(\tilde{x}[2]) &= \mathrm{Var}(\tilde{m} + \tilde{z_2}) = \mathrm{Var}(\tilde{m}) + \mathrm{Var}(\tilde{z_2}) \\
&= 10 + 1 = 11 \\
\mathrm{Cov}(\tilde{b}, \tilde{x}) &= \mathrm{E}[\tilde{b}\tilde{x}] - \mathrm{E}[\tilde{b}]\mathrm{E}[\tilde{x}] = \mathrm{E}[\tilde{b}\tilde{x}] \\
&= [\mathrm{E}[\tilde{x}[1]\tilde{b}]\ \mathrm{E}[\tilde{x}[2]\tilde{b}]]^T = [1\ 0]^T
\end{aligned}$$

This gives us:

$$\begin{aligned}
\Sigma_{\tilde{x}} &= \begin{bmatrix} \mathrm{Var}(\tilde{x}[1]) & \mathrm{Cov}(\tilde{x}[1], \tilde{x}[2]) \\ \mathrm{Cov}(\tilde{x}[2], \tilde{x}[1]) & \mathrm{Var}(\tilde{x}[2]) \end{bmatrix} = \begin{bmatrix} 12 & 10 \\ 10 & 11 \end{bmatrix} \\
\Sigma_{\tilde{x}}^{-1} &= \begin{bmatrix} \frac{11}{32} & -\frac{5}{16} \\ -\frac{5}{16} & -\frac{3}{8} \end{bmatrix} \\
\Sigma_{\tilde{b}\tilde{x}}^T \Sigma_{\tilde{x}}^{-1} \Sigma_{\tilde{b}\tilde{x}} &= [1\ 0] \begin{bmatrix} \frac{11}{32} \\ -\frac{5}{16} \end{bmatrix} = \frac{11}{32}
\end{aligned}$$

Hence MSE = $1 - \frac{11}{32} = \frac{21}{32} = 0.65$. The second estimator provides a better estimation of the heartbeat of the baby by jointly using the two microphones.

3. (Gaussian minimum MSE estimator) In this problem we derive the minimum MSE estimator of a random variable $\tilde{b}$ given another random variable $\tilde{a}$ when both are jointly Gaussian. To simplify matters we assume the mean of both random variables is zero.

(a) Let us define

$$\tilde{c} := \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\tilde{a}. \tag{3}$$

Consider the decomposition of $\tilde{b}$ into the sum of $\tilde{c}$ and $\tilde{b} - \tilde{c}$. Provide a geometric interpretation of this decomposition. This decomposition is the orthogonal projection of $\tilde{b}$ into a vector in the span of $\tilde{a}$: $\tilde{c}$ and a vector orthogonal to this hyperspace: $\tilde{b} - \tilde{c}$.

(b) Compute the conditional expectation of $\tilde{c}$ given $\tilde{a} = a$ for a fixed $a \in \mathbb{R}$.

$$\mathrm{E}[\tilde{c}|\tilde{a} = a] = \mathrm{E}[\frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\tilde{a}|\tilde{a} = a] = \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\tilde{a}$$

(c) Compute the conditional expectation of $\tilde{b} - \tilde{c}$ given $\tilde{a} = a$ for a fixed $a \in \mathbb{R}$. (Hint: Start by computing the covariance between $\tilde{b} - \tilde{c}$ and $\tilde{a}$.) First few results

$$\begin{aligned}
\mathrm{Cov}(\tilde{a}, \tilde{b}) &= \mathrm{E}[\tilde{a}\ \tilde{b}] - \mathrm{E}[\tilde{a}]\ \mathrm{E}[\tilde{b}] \\
&= \mathrm{E}[\tilde{a}\ \tilde{b}] - 0 = \mathrm{E}[\tilde{a}\ \tilde{b}] \\
\mathrm{Var}(\tilde{a}) &= \mathrm{E}[\tilde{a}^2] - \mathrm{E}[\tilde{a}]^2 = \mathrm{E}[\tilde{a}^2] \\
\mathrm{E}[\tilde{b} - \tilde{c}] &= \mathrm{E}[\tilde{b} - \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\tilde{a}] \\
&= \mathrm{E}[\tilde{b}] - \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\mathrm{E}[\tilde{a}] = 0
\end{aligned}$$

Hence

$$\mathrm{Cov}(\tilde{b} - \tilde{c}, \tilde{a}) = \mathrm{E}[\tilde{b}\ \tilde{a}] - \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\mathrm{E}[\tilde{a}^2] = \mathrm{Cov}(\tilde{a}, \tilde{b}) - \mathrm{Cov}(\tilde{a}, \tilde{b}) = 0$$

Thus $\tilde{b} - \tilde{c}$ and $\tilde{a}$ are uncorrelated and $\mathrm{E}[\tilde{b} - \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\tilde{a}|\tilde{a} = a] = \mathrm{E}[\tilde{b} - \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\tilde{a}] = \mathrm{E}[\tilde{b}] - \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\mathrm{E}[\tilde{a}] = 0$.

(d) Prove that the minimum MSE estimator of $\tilde{b}$ given $\tilde{a} = a$ for a fixed $a \in \mathbb{R}$ is linear.

Using the problem assumptions, and theorem 2.1 from our class, the minimum MSE estimator of $\tilde{b}$ given $\tilde{a} = a$ for a fixed $a \in \mathbb{R}$ is given by:

$$\begin{aligned}
\mathrm{E}[\tilde{b}|\tilde{a} = a] &= \mathrm{E}[\tilde{b} - \tilde{c} + \tilde{c}|\tilde{a} = a] \\
&= \mathrm{E}[\tilde{b} - \tilde{c}|\tilde{a} = a] + \mathrm{E}[\tilde{c}|\tilde{a} = a] \\
&= \frac{\mathrm{Cov}(\tilde{a}, \tilde{b})}{\mathrm{Var}(\tilde{a})}\tilde{a}
\end{aligned}$$

(e) What step of the proof fails for non-Gaussian random variables? Since $\tilde{a}$ and $\tilde{b}$ are gaussian random variables then $\tilde{a}$ and $\tilde{b} - \tilde{c}$ are also jointly gaussian. Furthermore $\mathrm{E}[\tilde{a}(\tilde{b} - \tilde{c})] = \mathrm{E}[\tilde{a}\tilde{b}] - \mathrm{E}[\tilde{a}\tilde{c}] = \mathrm{Cov}(\tilde{a}, \tilde{b}) - \mathrm{Cov}(\tilde{a}, \tilde{b}) = 0$. Thus $\tilde{a}$ and $\tilde{b} - \tilde{c}$ are uncorrelated and being gaussian are also independent. By linear combination of $\tilde{a}$, $\tilde{c}$ and $\tilde{b} - \tilde{c}$ are also independent. This allows in the first step, to decompose $\tilde{b}$ into two independent gaussian random variables: $\tilde{b} = \tilde{c} + \tilde{b} - \tilde{c}$.

4. (Oxford Dataset) In this problem, we will compute an estimator for rainfall in Oxford as a function of the maximum temperature. `oxford.zip` contains the support code for the problem and the dataset. `regression.py` within `oxford.zip` reads the dataset and splits it into train, validation and test sets. We parameterize our estimator for rainfall($y$) from maximum temperature($x$) as

$$f_a(x) = \begin{cases} w_1 x + b_1 & \text{if } x < a \\ w_2 x + b_2 & \text{if } x \geq a \end{cases}$$

$w_1, w_2, b_1$ and $b_2$ are estimated by minimizing the mean squared error on the training dataset.

(a) Complete `split_and_plot()` in `regression.py` to fit two different linear function for a given value of threshold $a$. The function will generate a plot of the fit overlaid on a scatter plot of the validation data. Report the plot generate by the function for different values of $a$ defined in `main()`. You are welcome to try other values of $a$, but please make sure that you report the plots generated for all values of $a$ defined in `main()`. In the function `split_and_plot()` in `regression.py`

(1) Split the training set for maximum temperatures into two data sets less or greater then temperature a ($\leq$ or $\geq$). We do the same for the rainfall datasets, splitting the rainfall dataset related to temperatures less or greater then temperature a ($\leq$ or $\geq$).

(2) Fit two linear models, one with the data related to temperatures less or equal to $a$ and the related rainfall data points, and one model with the training data corresponding to temperatures greater or equal to $a$ and related rainfall data points.

(3) We obtain two sets of prediction values on relevant points on grid using linear fit with points max_temp $\leq a$ and max_temps $\geq a$.

(4) We then compute the training mean squared errors for the two linear models by using the relevant training data points (model 1 using max_temp $\leq a$ and model 2 using max_temps $\geq a$) and comparing to the rainfall value split on $a$ with the same inequalities. The total training mse is a weighted average the two training mean squared errors:

$$\frac{n1 * \text{training mse 1} + n_2 * \text{training mse 2}}{n_1 + n_2}$$

$n_1$ : number of samples in data set used by linear model 1

$n_2$ : number of samples in data set used by linear model 2

(5) Following the same logic, we compute the total validation mse using the two linear model and the max. temperatures and rainfall validation data points.

(6) The lowest validation mse is $1134.86$ and it is obtained with two linear models using two datasets divided by a temperature $a = 20$.

We also tried various linear estimators: linear regression, stochastic gradient descent, and with $l2$ regularization: ridge and Bayesian ridge regressors, using different values of $a$ as split points to train the estimators. Out-of-the box the linear regression model had the best validation accuracy or the lowest validation error (see following plots for various $a$).

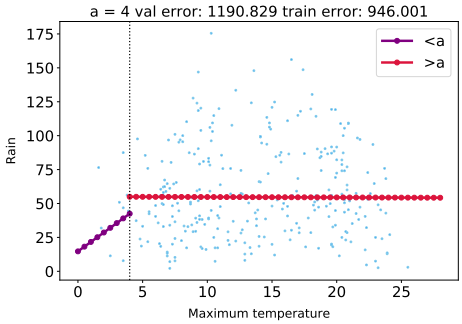


Figure 1: Linear regression of rain fall as response of max. temperatures with $a = 4$

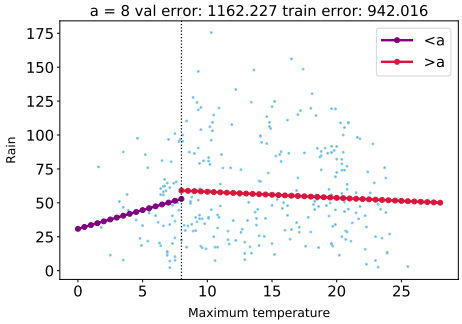


Figure 2: Linear regression of rain fall as response of max. temperatures with $a = 8$

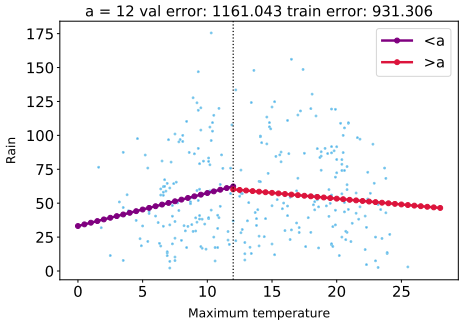


Figure 3: Linear regression of rain fall as response of max. temperatures with $a = 12$

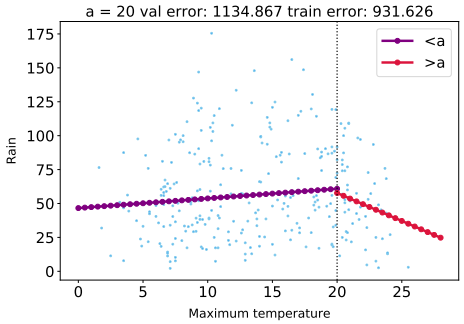


Figure 4: Linear regression of rain fall as response of max. temperatures with $a = 20$

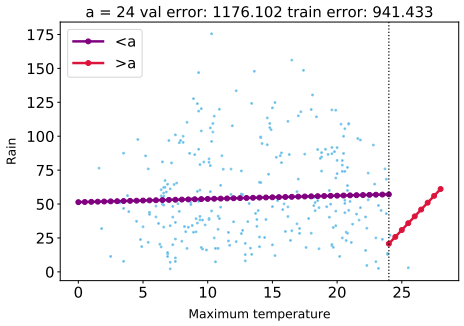


Figure 5: Linear regression of rain fall as response of max. temperatures with $a = 24$

(b) Choose the best estimator $f_{a'}(x)$ according to the validation error. Fill in the rest of `main()` function to fit a single linear estimator on the entire dataset. Compare the fit and error values of $f_{a'}(x)$ with the single linear estimator fit on the training set on the held out test set. Report the plot generated by this section.

Having identified the best value $a$ ($a = 20$) to split the training data, we then train two linear models using this value of $a$ by splitting the entire training dataset between max.temperatures $\leq a$ and max. temperatures $\geq a$ and fitting with the split on the rainfalls using the same value of $a$. We then train a linear model on the entire training set using the train maximum temperatures and train rainfall data. We observer that our best estimator $f_{a'}(x)$ has better accuracy or equivalently lower mean squared error: $1487.4$ vs. $1680$ for the single linear model, Splitting the data into two datasets allowed to have better predictions improving the acuray by $11\%$.

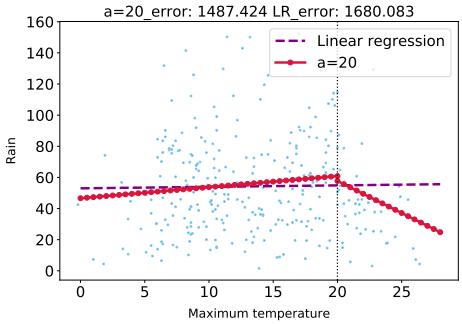


Figure 6: Comparison between a single linear estimator fitted on the whole dataset and the the best estimator $f'_a$

We do not require you to include your code in the report. You can choose to include it or not include it.