CS-E3210- Machine Learning Basic Principles
Home Assignment 2 - "Regression"

Your solutions to the following problems should be submitted as one single pdf which does not contain any personal information (student ID or name). The only rule for the layout of your submission is that for each problem there has to be exactly one separate page containing the answer to the problem. You are welcome to use the LaTeX-file underlying this pdf, available under `https://version.aalto.fi/gitlab/junga1/MLBP2017Public`, and fill in your solutions there.

Problem 1: "Plain Vanilla" Linear Regression

Consider a dataset $\mathbb{X}$ which is constituted of $N=10$ webcam snapshots with filename "MontBlanc*i*.png", $i = 1,\ldots,N$, available in the folder "Webcam" at `https://version.aalto.fi/gitlab/junga1/MLBP2017Public`. Determine for each snapshot the feature vector $\mathbf{x}^{(i)} = (x_{\rm g}^{(i)},1)^T \in \mathcal{X}(=\mathbb{R}^2)$ with the normalized (by the number of image pixels) greenness $x_{\rm g}^{(i)}$. Moreover, determine for each snapshot the label $y^{(i)} \in \mathcal{Y}(=\mathbb{R})$ given by the duration (in minutes) after 07:00 am, at which the picture has been taken. We want to find (learn) a predictor $h(\cdot): \mathcal{X} \rightarrow \mathcal{Y}$ which allows to predict the value of $y^{(i)}$ directly from the value of the feature $x_{\rm g}^{(i)}$. To this end we consider only predictors belonging to the hypothesis space $\mathcal{H} = \{h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} \text{ for some } \mathbf{w} \in \mathbb{R}^2\}$. The quality of a particular predictor is measured by the mean squared error

$$\mathcal{E}(h(\cdot)|\mathbb{X}) := \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - h(\mathbf{x}^{(i)}))^2 = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2. \tag{1}$$

Note that the mean squared error is nothing but the empirical risk obtained when using the squared error loss $L((\mathbf{x},y),h(\cdot)) = (y-h(\mathbf{x}))^2$ (cf. Lecture 2).

The optimal predictor $h_{\rm opt}(\cdot)$ is then

$$h_{\rm opt}(\cdot) = \operatorname*{argmin}_{h(\cdot)\in\mathcal{H}} \mathcal{E}(h(\cdot)|\mathbb{X}). \tag{2}$$

We can rewrite this optimization problem in a fully equivalent manner in terms of the weight $\mathbf{w}$ representing a particular predictor $h^{(\mathbf{w})}(\cdot) \in \mathcal{H}$ as

$$\mathbf{w}_{\rm opt} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^2} \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2. \tag{3}$$

As can be verified easily, the optimal predictor $h_{\rm opt}(\cdot)$ (cf. (2)) is obtained as $h_{\rm opt}(\cdot) = h^{(\mathbf{w}_{\rm opt})}(\cdot)$ with the optimal weight vector $\mathbf{w}_{\rm opt}$ (cf. (3)).

Can you find a closed-form expression for the optimal weight $\mathbf{w}_{\rm opt}$ (cf. (3)) in terms of the vectors $\mathbf{x} = (x_{\rm g}^{(1)},\ldots,x_{\rm g}^{(N)})^T \in \mathbb{R}^N$, and $\mathbf{y} = (y^{(1)},\ldots,y^{(N)})^T \in \mathbb{R}^N$?

Answer. Using $\mathbf{X} = [\mathbf{1}_N, \mathbf{x}] = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ x_g^{(1)} & x_g^{(2)} & \ldots & x_g^{(N)} \end{bmatrix}^T$, with the all-ones vector $\mathbf{1}_N = (1,1,\ldots,1)^T \in \mathbb{R}^N$, the mean square error $\mathcal{E}(h(\cdot)|\mathbb{X})$ in (1) can be written as $\mathcal{E}(h(\cdot)|\mathbb{X}) = \frac{1}{N}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2$. For a given data matrix $\mathbf{X}$ and vector $\mathbf{y}$, the mean square error depends only on the weight vector $\mathbf{w}$ and thus defines a function $f(\mathbf{w}) := \mathcal{E}(h(\cdot)|\mathbb{X}) = \frac{1}{N}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2$. We would like to find the particular choice $\mathbf{w}_{\rm opt}$ which minimizes this function, i.e., $f(\mathbf{w}_{\rm opt}) = \min_{\mathbf{w}\in\mathbb{R}^2} f(\mathbf{w})$. Since the function $f(\mathbf{w}) = \frac{1}{N}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2$ is differentiable and convex, a necessary and sufficient condition for a vector $\mathbf{w}_{\rm opt}$ to yield a minimum is the zero-gradient condition (see p. 140 of the book `https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`):

$$f(\mathbf{w}_{\rm opt}) = \min_{\mathbf{w}\in\mathbb{R}^2} f(\mathbf{w}) \text{ if and only if } \nabla f(\mathbf{w}_{\rm opt}) = \mathbf{0}. \tag{4}$$

The gradient $\nabla f(\mathbf{w})$ of the function $f(\mathbf{w}) = \frac{1}{N}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2$ is obtained as (see p. 641 of the book `https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`)

$$\nabla f(\mathbf{w}) = \nabla \frac{1}{N}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2 = (-2/N)\mathbf{X}^T(\mathbf{y}-\mathbf{X}\mathbf{w}). \tag{5}$$

Combining (4) with (5), we arrive at the following characterization of $\mathbf{w}_{\rm opt}$:

$$f(\mathbf{w}_{\rm opt}) = \min_{\mathbf{w}\in\mathbb{R}^2} f(\mathbf{w}) \text{ if and only if } \mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\mathbf{w}_{\rm opt}. \tag{6}$$

If the matrix $\mathbf{X}^T\mathbf{X}$ is invertible (which is the case when $\mathbf{x}$ does not contain identical entries, i.e., $\mathbf{x}$ is linearly independent from the all-ones vector $\mathbf{1}_N$), we obtain from (6) the following closed-form expression for the optimal weight vector:

$$\mathbf{w}_{\rm opt} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{7}$$

Problem 2: "Plain Vanilla" Linear Regression - Figure

Reconsider the setup of Problem 1 and generate a plot with horizontal (vertical) axis representing greenness $x_{\rm g}$ (label y), which depicts the optimal predictor $h_{\rm opt}(\cdot)$ (cf. (2)) and also contains the data points $(x_{\rm g}^{(i)}, y^{(i)})$ for $i = 1, \ldots, N$. Do you consider it feasible to predict the daytime accurately from the greenness?

Answer. In Fig. 1, we depict the optimal predictor $h_{\rm opt}(\cdot)$ (cf. (2)) along with the data points $(x_{\rm g}^{(i)}, y^{(i)})$ for $i = 1, \ldots, N$.

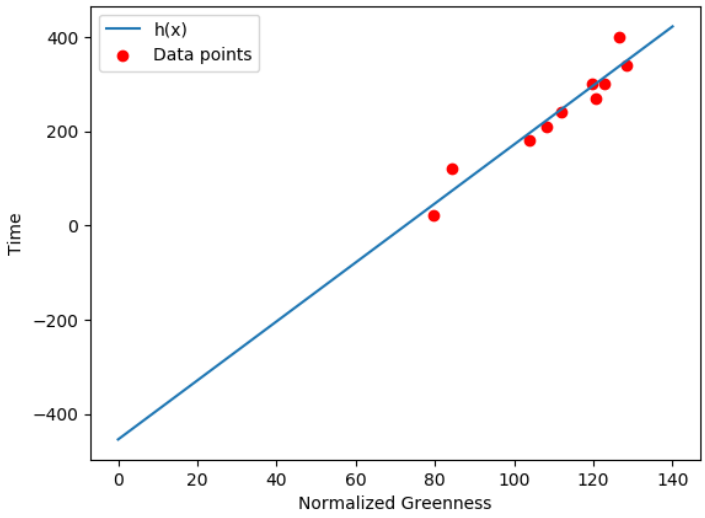

Figure 1: Optimal predictor $h_{\rm opt}(\cdot) = \mathbf{w}_{\rm opt}^T \mathbf{x}$ and the data points $(x_{\rm g}^{(i)}, y^{(i)})$.

According to Fig. 1, it seems feasible to roughly predict the daytime from the greenness of the snapshot.

Problem 3: Regularized Linear Regression

We consider again the regression problem of Problem 1, i.e., predicting the daytime of a webcam snapshot based on the feature vector $(x_{\rm g}, 1)^T$. The prediction is of the form $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ with some weight vector $\mathbf{w} \in \mathbb{R}^2$. Assume that we only have snapshots which are taken within 7 hours after 07:00 am, i.e.,the value of the label y cannot exceed 420. Therefore, it makes sense to somehow constraint the norm of the weight vector $\mathbf{w}$ to exclude unreasonable predictions. To this end, we augment the mean squared error (1) with the "regularization term" $\lambda\|\mathbf{w}\|^2$ which penalizes "atypical" values for the weight vector. The optimal predictor $h_{\rm opt}(\cdot)$ using this regularization term is then given by

$$h_{\rm opt,r}(\cdot) = \operatorname*{argmin}_{h^{(\mathbf{w})}(\cdot)\in\mathcal{H}} \big(\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) + \lambda\|\mathbf{w}\|^2\big). \tag{8}$$

Again, we can rewrite this optimization problem in a fully equivalent manner in terms of the weight $\mathbf{w}$ representing a particular predictor $h^{(\mathbf{w})}(\cdot) \in \mathcal{H}$ as

$$\mathbf{w}_{\rm opt,r} = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^2} \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2 + \lambda\|\mathbf{w}\|^2. \tag{9}$$

As can be verified easily, the optimal predictor $h_{\rm opt,r}(\cdot) \in \mathcal{H}$ solving (2) is obtained as $h_{\rm opt,r}(\cdot) = h^{(\mathbf{w}_{\rm opt,r})}(\cdot)$ with the optimal weight vector $\mathbf{w}_{\rm opt,r}$ which is the solution of (3). Can you find a closed-form solution for the optimal weight $\mathbf{w}_{\rm opt,r}$ (cf. (3)) in terms of the vectors $\mathbf{x} = (x_{\rm g}^{(1)}, \ldots, x_{\rm g}^{(N)})^T \in \mathbb{R}^N$, and $\mathbf{y} = (y^{(1)}, \ldots, y^{(N)})^T \in \mathbb{R}^N$ and λ?

Answer:

In order to find the optimal weight $\mathbf{w}_{\rm opt,r}$, we can proceed along very similar lines as in Problem 1. The only difference to Problem 1 is that we now aim at minimizing the regularized empirical risk $\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) + \lambda\|\mathbf{w}\|^2$ instead of the empiricial risk $\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X})$, which was minimized in Problem 1. Thus, we now have to minimize the objective function $f_r(\mathbf{w}) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|_2^2$ instead of $f(\mathbf{w}) = \frac{1}{N}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$ (as in Problem 1). The function $f_r(\mathbf{w})$ is again convex and differentiable, so we can use the zero-gradient condition to characterize the optimal weight $\mathbf{w}_{\rm opt,r}$. Moreover, the gradient of $f_r(\mathbf{w})$ can be obtained using a similar calculation as in Problem 1, yielding $\nabla f_r(\mathbf{w}) = (-2/N)\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda\mathbf{w}$. We obtain the following characterization of the optimal weight $\mathbf{w}_{\rm opt,r}$:

$$f_r(\mathbf{w}_{\rm opt,r}) = \min_{\mathbf{w}\in\mathbb{R}^2} f_r(\mathbf{w}) \text{ if and only if } (1/N)\mathbf{X}^T\mathbf{y} = ((1/N)\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w}_{\rm opt,r}. \tag{10}$$

Since for any positive $\lambda > 0$, the matrix $((1/N)\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$ is guaranteed to be invertible, we obtain finally

$$\mathbf{w}_{\rm opt,r} = \big(\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I}\big)^{-1}\mathbf{X}^T\mathbf{y}. \tag{11}$$

Problem 4: Regularized Linear Regression - Figure

Reconsider the setup of Problem 3 and generate a plot with horizontal (vertical) axis representing greenness x_{g} (label y) which contains the data points $(x_{\mathrm{g}}^{(i)}, y^{(i)})$, for $i = 1, \ldots, N$, and depicts the optimal predictor $h_{\mathrm{opt,r}}(\cdot)$ (cf. (8)) for the two particular choices $\lambda = 2$ and $\lambda = 5$. Which choice for λ seems to be better for the given task?

Answer: In Fig. 2 we have the required plots. We can intuitively tell by comparing Fig. 2 with Fig. 1 at the plots that regularization using this particular values for λ had a detrimental effect on the learnt weight vector.

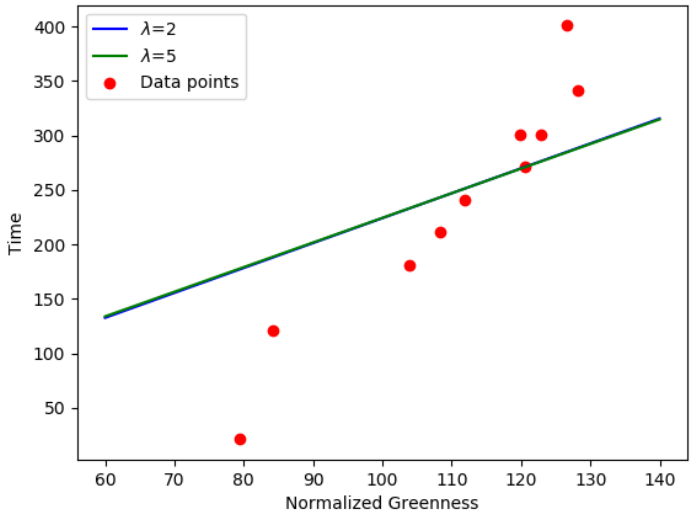

Figure 2: Predictors $h(\mathbf{x}) = \mathbf{w}_{(\mathrm{opt,r})}^T \mathbf{x}$ using the optimal vectors obtained for regularization parameter $\lambda = 2$ and $\lambda = 4$. We also show the data points.

The predictors obtained using reguarlized linear regression using regularaization parameter $\lambda = 2, 5$ are almost indistinguishable and overall seem to be performing worse compared to the predictor obtained without regularization (in Problem 1 and 2).

Problem 5: Gradient Descent for Linear Regression

Consider the same dataset as in Problem 1, i.e., the set of $N = 10$ webcam snapshots which are labeled by the daytime $y^{(i)}$ when the image has been taken. As in Problem 1, we are interested in predicting the daytime directly from the image. However, by contrast to Problem 1 where we only used the greenness $x_{\mathrm{g}}^{(i)}$ of the i-th image, we know use the green intensity values for the upper-left area consisting of 100×100 pixels, which we stack into the feature vector $\mathbf{x}^{(i)} \in \mathbb{R}^d$. What is the length d of the feature vector $\mathbf{x}^{(i)}$ here? Based on the feature vector, we predict the daytime by a predictor of the form $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ with some weight vector $\mathbf{w} \in \mathbb{R}^d$. The optimal predictor is obtained by solving an empirical risk minimization problem of the form (2), or directly in terms of the weight vector, (3). This minimization problems can be solved by a simple but powerful iterative method known as gradient descent (GD):

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha \nabla f(\mathbf{w}^{(k)}) \tag{12}$$

with some positive step size $\alpha > 0$ and the mean-squared error cost function (cf. (1))

$$f(\mathbf{w}) := \mathcal{E}(h^{(\mathbf{w})}|\mathbb{X}) \overset{(1)}{=} \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2.$$

In order to implement the GD iterations (12), we need to compute the gradient $\nabla f(\mathbf{w})$. Can you find a simple closed-form expression for the gradient of $f(\mathbf{w})$ at a particular weight vector $\mathbf{w}$?

The performance of GD depends crucially on the particular value chosen for the step size α in (12). Try out different choices for the step size α and, for each choice plot the evolution of the empirical risk $\mathcal{E}(h^{(\mathbf{w}^{(k)})}|\mathbb{X})$ as a function of iteration number k into one single figure. Use the initialization $\mathbf{w}^{(0)} = \mathbf{0}$ for the GD iterations for each run.

Another crucial issue when using GD is the question of when to stop iterating (12). Can you state a few stopping criteria that indicate when it would be reasonable to stop iterating (12)?

Answer: The dimension is $d = 100^2 = 10000$. The gradient is obtained as

$$\nabla f(\mathbf{w}) = \frac{-2}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}). \tag{13}$$

For the stopping criterion we might use a fixed number of iterations, which requires to have some understanding ("convergence analysis") of how fast gradient descent converges to the optimum. Another option is to monitor the relative decrease of the objective value $f(\mathbf{w}$, i.e., to stop iterating when $\left|\frac{f(\mathbf{w}^{(k+1)}) - f(\mathbf{w}^{(k)})}{f(\mathbf{w}^{(k)})}\right|$ is below a suitably chosen (small) threshold.

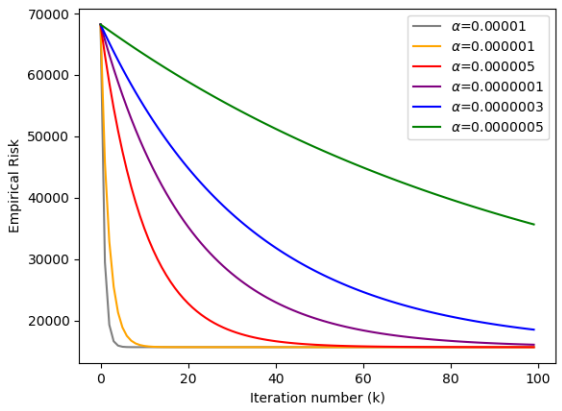

Figure 3: The convergence of the empirical risk with different step sizes α.

Problem 6: Gradient Descent for Regularized Linear Regression

Redo Problem 5 by using regularized linear regression (cf. Problem 3) instead of linear regression.
Answer: The dimension is $d = 100^2 = 10000$. The gradient is obtained as

$$\nabla f(\mathbf{w}) = 2\lambda \mathbf{w} + \frac{-2}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)}). \tag{14}$$

For the stopping criterion we might use a fixed number of iterations, which requires to have some understanding ("convergence analysis") of how fast gradient descent converges to the optimum. Another option is to monitor the relative decrease of the objective value $f(\mathbf{w})$, i.e., to stop iterating when $\left| \frac{f(\mathbf{w}^{(k+1)}) - f(\mathbf{w}^{(k)})}{f(\mathbf{w}^{(k)})} \right|$ is below a suitably chosen (small) threshold.

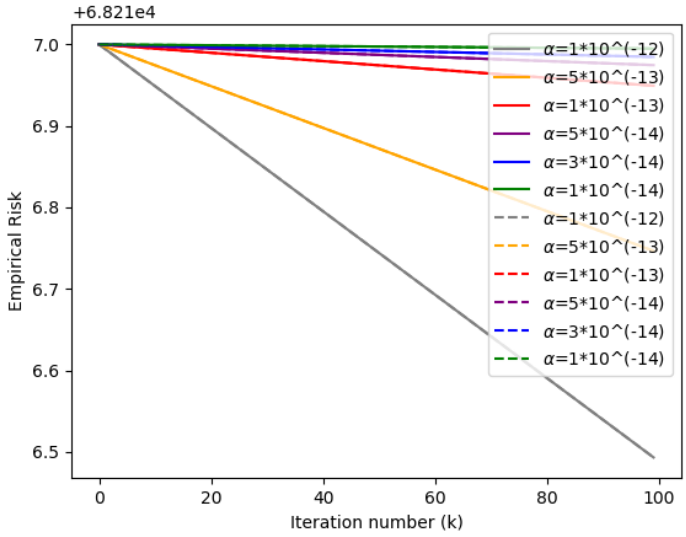

Figure 4: Plot of the empirical risk obtained when choosing the weight vector minimizing the regularized empirical risk with $\lambda = 2$ (dashed lines) and $\lambda = 5$ (solid lines). Each of the lines are obtained for different choices for the step-size α. Note that the curves for the same value of α are on top of each other, i.e., there is virtually no difference the results obtained from using $\lambda = 2$ or $\lambda = 5$.

Problem 7: Kernel Regression

Consider the data set of Problem 1, i.e., the set of $N = 10$ webcam snapshots. Let us now represent each webcam snapshot by the single feature $x^{(i)} = x_{\rm g}^{(i)}$, i.e., the total greenness of the ith snapshot. We aim at predicting the daytime $y^{(i)}$ based solely on the greenness. In contrast to Problem 1 and Problem 2 we will now use a different hypothesis space of predictors. In particular, we only consider predictors out of the hypothesis space

$$\mathcal{H} = \left\{ h^{(\sigma)}(\cdot) : \mathbb{R} \to \mathbb{R} : h^{(\sigma)}(x) = \sum_{i=1}^{N} y^{(i)} \frac{K_\sigma(x, x^{(i)})}{\sum_{l=1}^{N} K_\sigma(x, x^{(l)})} \right\} \tag{15}$$

with the "kernel"

$$K_\sigma(x, x^{(i)}) = \exp\left(-\frac{1}{2} \frac{(x - x^{(i)})^2}{\sigma^2} \right). \tag{16}$$

Try out predicting the daytime $y^{(i)}$ using the greenness $x_{\rm g}^{(i)}$ using a predictor $h^{(\sigma)}(\cdot) \in \mathcal{H}$ using the choices $\sigma \in \{1, 5, 10\}$. Generate a plot with horizontal (vertical) axis representing greenness $x_{\rm g}$ (label y), which depicts the predictor $h^{(\sigma)}(\cdot)$ for $\sigma \in \{1, 5, 10\}$ and also contains the data points $(x_{\rm g}^{(i)}, y^{(i)})$. Which choice for σ achieves the lowest mean squared error $\mathcal{E}(h^{(\sigma)}|\mathbb{X})$ (cf. (1)) ?

Answer: In Fig. 5, we depict the predictor $h^{(\sigma)}(\cdot)$, which maps the input feature $x_{\rm g}$ to the predicted daytime $h(x_{\rm g})$, obtained for $\sigma \in \{1, 5, 10\}$.

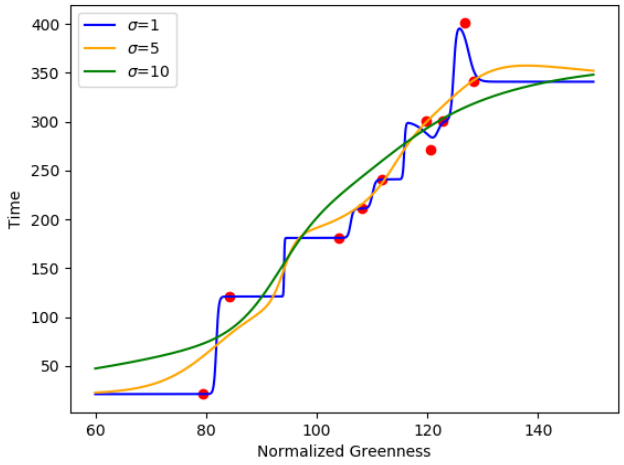

Figure 5: Image presenting the scatter plot and the linear regression plot of time in relation to greenness in the snapshots provided.

The empirical risk obtained by the predictor $h^{(\sigma)}(\cdot)$ obtained for $\sigma \in \{1, 5, 10\}$ is listed in Table 1.

σ value	Mean Squared Error	Root Mean Squared Error	Normalized Root Mean Squared Error
$\sigma = 1$	69.988	8.366	0.0350
$\sigma = 5$	908.282	30.138	0.1261
$\sigma = 10$	1584.557	39.806	0.1666

Table 1: Empirical risk for the three different values of variable σ.

The lowest mean squared error is obtained for the choice $\sigma = 1$.

Problem 8: Linear Regression using Feature Maps

Consider a regression problem, where we aim at predicting the value of a real-valued label or target or output variable $y \in \mathbb{R}$ of a data point based on a single feature $x \in \mathbb{R}$ of this data point. We assume that there is some true underlying functional relationship between feature x and output y, i.e., $y = h^*(x)$ with some unknown function (hypothesis). All we know about this true underlying functional relationship is that

$$h^*(x) = 0 \text{ for any } x \notin [0,10], \text{ and } |h^*(x') - h^*(x'')| \leq 10^{-3}|x' - x''| \text{ for any } x', x'' \in [0,10]. \tag{17}$$

We apply then a feature map $\boldsymbol{\phi} : \mathbb{R} \to \mathbb{R}^n$, with some suitable chosen dimension n, which transforms the original feature x into a modified feature vector $\boldsymbol{\phi}(x) = (\phi_1(x), \ldots, \phi_n(x))^T$. We use the transformed features $\boldsymbol{\phi}(x)$ to predict the label y using the predictor $h^{(\mathbf{w})}(x) = \mathbf{w}^T\boldsymbol{\phi}(x)$ with some weight vector $\mathbf{w} \in \mathbb{R}^n$. Note that the so defined predictor $h^{(\mathbf{w})}$ is linear only w.r.t. the high-dimensional features $\boldsymbol{\phi}(x)$, but typically a non-linear function of the original feature x. Is there a feature map $\boldsymbol{\phi}$ which allows to approximate the true hypothesis $h^*(\cdot)$ (which satisfies (17)) by some predictor $h^{(\mathbf{w}_0)}(x) = \mathbf{w}_0^T\boldsymbol{\phi}(x)$ with a suitably chosen weight $\mathbf{w}_0$? In particular, is there a feature map $\boldsymbol{\phi}$ and weight vector $\mathbf{w}_0 \in \mathbb{R}^n$ such that $|h^{(\mathbf{w}_0)}(x) - h^*(x)| \leq 10^{-3}$ for all $x \in \mathbb{R}$?

Answer:

Yes there is. Let us partition the interval $[0,10]$ into the non-overlapping subintervals $\mathcal{B}_1 = [0,2], \mathcal{B}_2 = (2,4], \mathcal{B}_3 = (4,6], \mathcal{B}_4 = (6,8], \mathcal{B}_4 = (8,10]$, i.e., $[0,10] = \mathcal{B}_1 \cup \mathcal{B}_2 \ldots \cup \mathcal{B}_5$. Using this partition, we construct the feature map $\boldsymbol{\phi}(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_5(x)) \in \mathbb{R}^5$ with components $\phi_j(x) = \mathcal{I}(x \in \mathcal{B}_j)$. For an arbitrary hypothesis $h^*(\cdot)$ which satisfies (17), we construct the weight vector

$$\mathbf{w}_0 = (h^*(1), h^*(3), h^*(5), h^*(7), h^*(9))^T \in \mathbb{R}^5. \tag{18}$$

Let us show that $h^{(\mathbf{w}_0)}(x) = \mathbf{w}_0^T\boldsymbol{\phi}(x) = \sum_{j=1}^5 h^*(2j-1)\phi_j(x)$ satisfies the desired condition, i.e.,

$$|h^{(\mathbf{w}_0)}(x) - h^*(x)| \leq 10^{-3}, \text{ for all } x \in \mathbb{R}. \tag{19}$$

We will verify (19) by considering separately the two complementary cases: $x \notin [0,10]$ and $x \in [0,10]$.

- when $x \notin [0,10]$: by (17), $h^*(x) = 0$; since $h^{(\mathbf{w}_0)}(x) = \mathbf{w}_0^T\boldsymbol{\phi}(x) = \mathbf{w}_0^T(0,0,0,0,0)^T = 0$, we have $|h^{(\mathbf{w}_0)}(x) - h^*(x)| = |0-0| = 0 \leq 10^{-3}$.
- when $x \in [0,10]$: consider the subinterval $\mathcal{B}_j$ which contains x, i.e., $x \in \mathcal{B}_j$. Let $r = \mathcal{B}_j \cap \{1,3,5,7,9\}$ and note that $|x - r| \leq 1$ as well as $\mathbf{w}_0^T\boldsymbol{\phi}(x) = h^*(r)$. Thus,

$$|h^{(\mathbf{w}_0)}(x) - h^*(x)| = |\mathbf{w}_0^T\boldsymbol{\phi}(x) - h^*(x)| = |h^*(r) - h^*(x)| \overset{(17)}{\leq} 10^{-3}|r - x| \leq 10^{-3}. \tag{20}$$

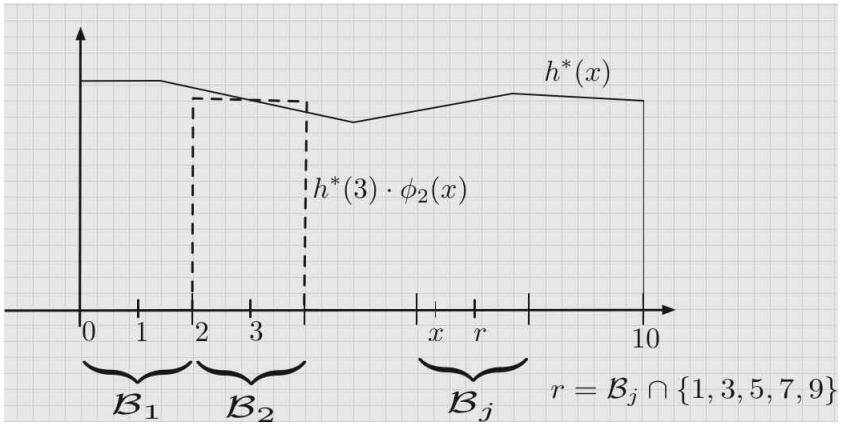

Figure 6: True hypothesis $h^*(x)$ is non-zero over the interval $[0,10]$ which is partitioned into subintervals $\mathcal{B}_1, \ldots$. For any point $x \in [0,10]$, we can find some $r \in \{1,3,5,7,9\}$ such that $|x - r| \leq 1$.