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

Answer.

$$\mathbf{X} = [\mathbf{x} \quad \mathbf{1}]$$

To minimize empirical risk $\mathcal{E}(h(\cdot)|\mathbb{X})$,we can directly solve where its gradients are 0:

$$\nabla_w \mathcal{E}(h(\cdot)|\mathbb{X}) = 0 \tag{1}$$

$$\frac{1}{N}\nabla_w \|\mathbf{Xw} - \mathbf{y}\|_{\mathbf{2}}^{\mathbf{2}} = \mathbf{0} \tag{2}$$

$$\nabla_w (\mathbf{Xw} - \mathbf{y})^{\mathbf{T}} (\mathbf{Xw} - \mathbf{y}) = \mathbf{0} \tag{3}$$

$$\nabla_w (\mathbf{w}^T \mathbf{X}^{\mathbf{T}} \mathbf{Xw} - \mathbf{2w}^{\mathbf{T}} \mathbf{X}^{\mathbf{T}} \mathbf{y} + \mathbf{y}^{\mathbf{T}} \mathbf{y}) = \mathbf{0} \tag{4}$$

$$2\mathbf{X}^{\mathbf{T}} \mathbf{Xw} - \mathbf{2X}^{\mathbf{T}} \mathbf{y} = \mathbf{0} \tag{5}$$

$$\mathbf{w}_{\text{opt}} = (\mathbf{X}^{\mathbf{T}} \mathbf{X})^{-\mathbf{1}} \mathbf{X}^{\mathbf{T}} \mathbf{y} \tag{6}$$

(Notice that $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ should be invertible)

Problem 2: "Plain Vanilla" Linear Regression - Figure

Answer.

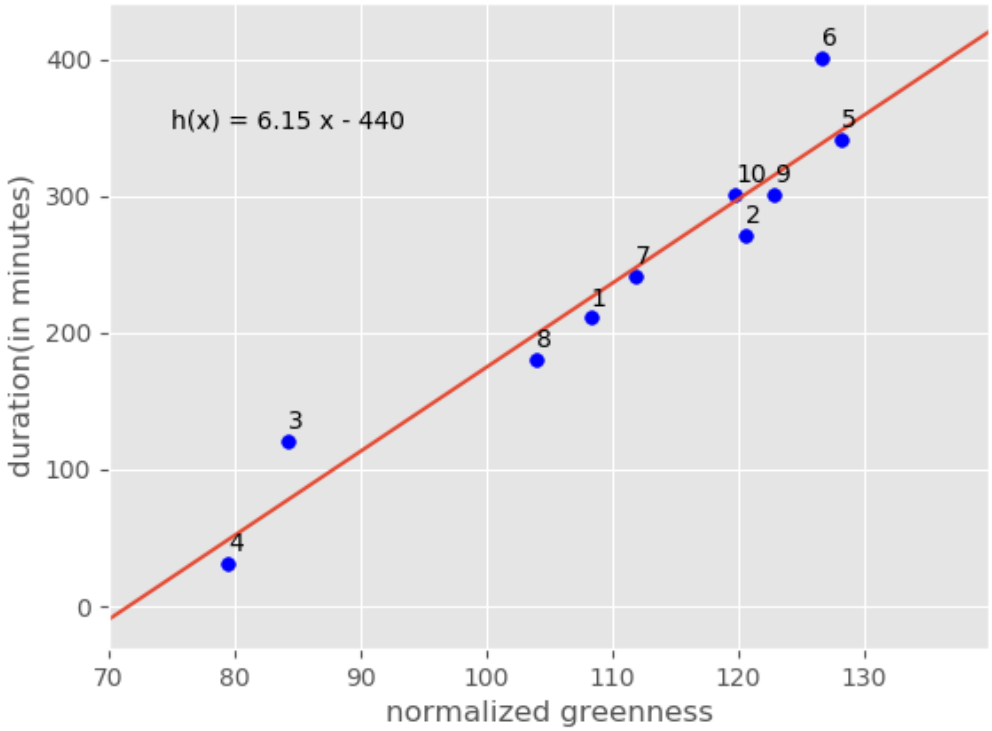

Figure 1: Plot of the optimal predictor and the data points

$$\mathbf{w}_{\text{opt}} \approx \begin{bmatrix} 6.15 \\ -440 \end{bmatrix} \tag{7}$$

The empirical risk:

$$\begin{aligned} \mathcal{E}(h(\cdot)|\mathbb{X}) &= \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - h(\mathbf{x}^{(i)}))^2 \\ &\approx 783 \end{aligned} \tag{8}$$

The empirical risk is not small enough, so it is not feasible to predict the daytime accurately from the greenness.

Problem 3: Regularized Linear Regression

Answer:

$$\mathbf{X} = [\mathbf{x} \quad \mathbf{1}]$$

$$\nabla_w \big(\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) + \lambda \|\mathbf{w}\|^2\big) = 0 \tag{9}$$

$$\nabla_w \big(\frac{1}{N}\|\mathbf{Xw} - \mathbf{y}\|_{\mathbf{2}}^{\mathbf{2}} + \lambda \|\mathbf{w}\|^{\mathbf{2}}\big) = \mathbf{0} \tag{10}$$

$$\nabla_w \big(\frac{1}{N}(\mathbf{Xw} - \mathbf{y})^{\mathbf{T}}(\mathbf{Xw} - \mathbf{y}) + \lambda \mathbf{w}^{\mathbf{T}}\mathbf{w}\big) = \mathbf{0} \tag{11}$$

$$\frac{1}{N}\nabla_w (\mathbf{Xw} - \mathbf{y})^{\mathbf{T}}(\mathbf{Xw} - \mathbf{y}) + \nabla_{\mathbf{w}}(\lambda \mathbf{w}^{\mathbf{T}}\mathbf{w}) = \mathbf{0} \tag{12}$$

From the Problem 1,

$$\nabla_w (\mathbf{Xw} - \mathbf{y})^{\mathbf{T}}(\mathbf{Xw} - \mathbf{y}) = \mathbf{2X^T Xw} - \mathbf{2X^T y} \tag{13}$$

$$\begin{aligned} &\frac{1}{N}(2\mathbf{X^T Xw} - \mathbf{2X^T y}) + \nabla_{\mathbf{w}}(\lambda \mathbf{w}^{\mathbf{T}}\mathbf{w}) \\ =&\frac{1}{N}(2\mathbf{X^T Xw} - \mathbf{2X^T y}) + \mathbf{2}\lambda \mathbf{w} \\ =&0 \end{aligned} \tag{14}$$

$$\mathbf{X^T Xw} - \mathbf{X^T y} + \mathbf{N}\lambda \mathbf{Iw} = \mathbf{0} \tag{15}$$

$$\mathbf{w} = (\mathbf{X^T X} + \mathbf{N}\lambda \mathbf{I})^{-\mathbf{1}}\mathbf{X^T y} \tag{16}$$

Problem 4: Regularized Linear Regression - Figure

Answer:

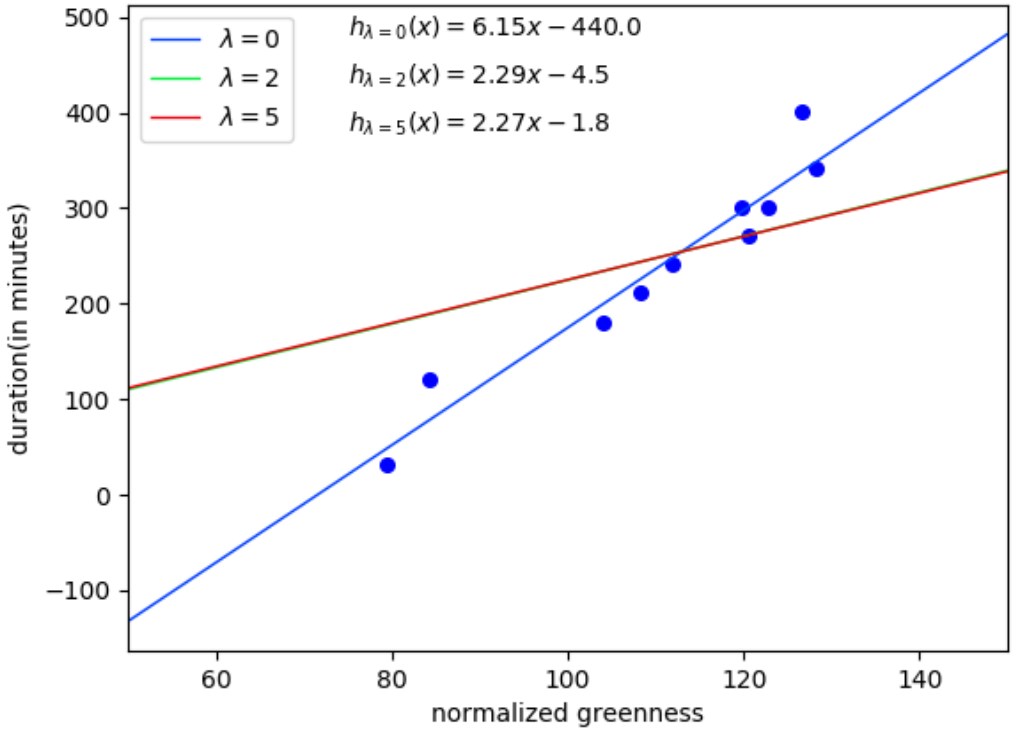

Figure 2: Plot of the regularized linear regression

$(\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) + \lambda\|\mathbf{w}\|^2)_{\lambda=2} \approx 4804$
$(\mathcal{E}(h^{(\mathbf{w})}(\cdot)|\mathbb{X}) + \lambda\|\mathbf{w}\|^2)_{\lambda=5} \approx 4844$

So $\lambda = 2$ is a better choice.

Problem 5: Gradient Descent for Linear Regression

Answer:
the length d of the feature vector $\mathbf{x}(i)$ is $100 \times 100 = 10000$.
(If taking into account dummy features 1, the length of the feature vector is 10001)

$$\mathbf{X} = [\mathbf{x} \quad \mathbf{1}]$$

$$\begin{aligned} \nabla_{\mathbf{w}} f(\mathbf{w}) &= \nabla_{\mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 \\ &= -\frac{2}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} (y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)}) \\ &= \frac{2}{N} (\mathbf{X^T X w} - \mathbf{X^T y}) \end{aligned} \tag{17}$$

We can also get the closed-form expression from the Problem 1.

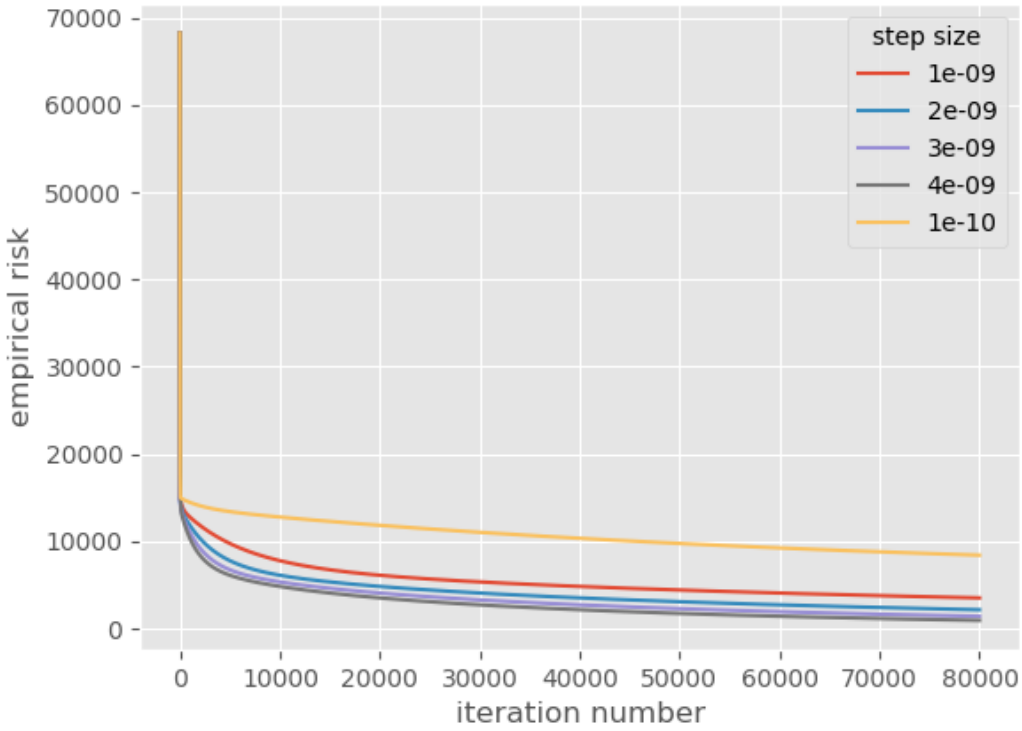

Figure 3: Gradient Descent for Linear Regression

Stopping criteria:
1. Identify when the error (emprical risk, etc.) is small enough to stop.
2. Stop when the gradient is small enough (the difference no longer decreases or decreases too slowly).
3. Limit the maximum amount of time spent iterating.

Problem 6: Gradient Descent for Regularized Linear Regression

Answer:
From the Problem 3,

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \frac{2}{N}(\mathbf{X^T X w} - \mathbf{X^T y}) + \mathbf{2}\lambda\mathbf{w} \tag{18}$$

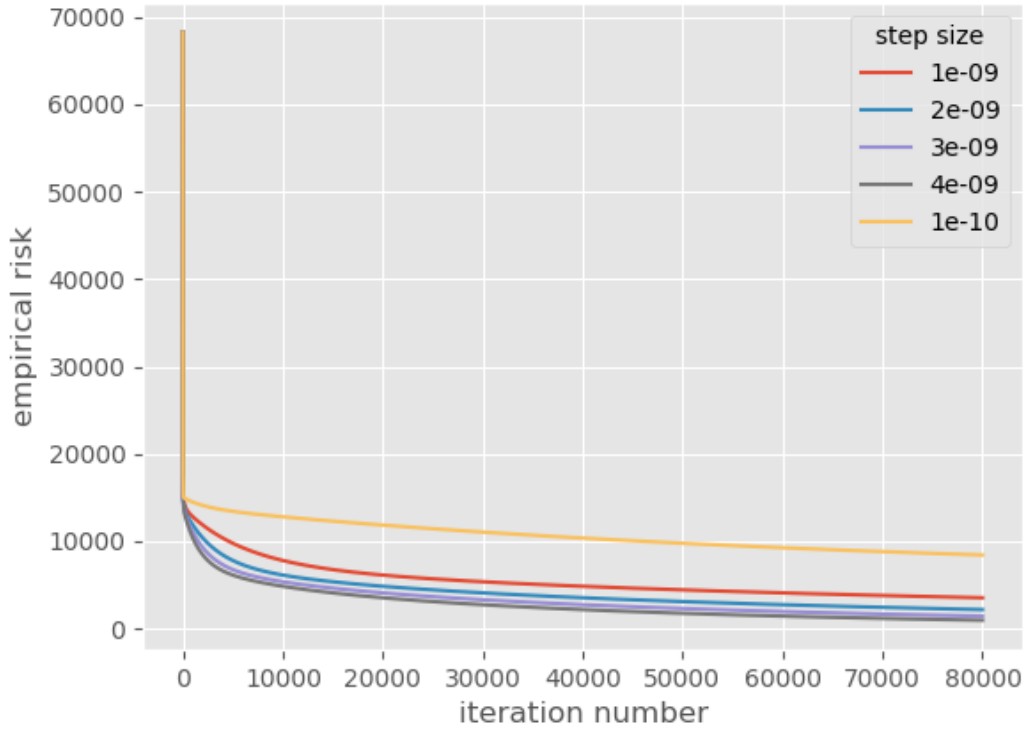

Figure 4: Gradient Descent for Regularized Linear Regression

After iterations, the converged empirical risk is bigger than that of linear regression (it is not obvious in the figure).

Problem 7: Kernel Regression

Answer:

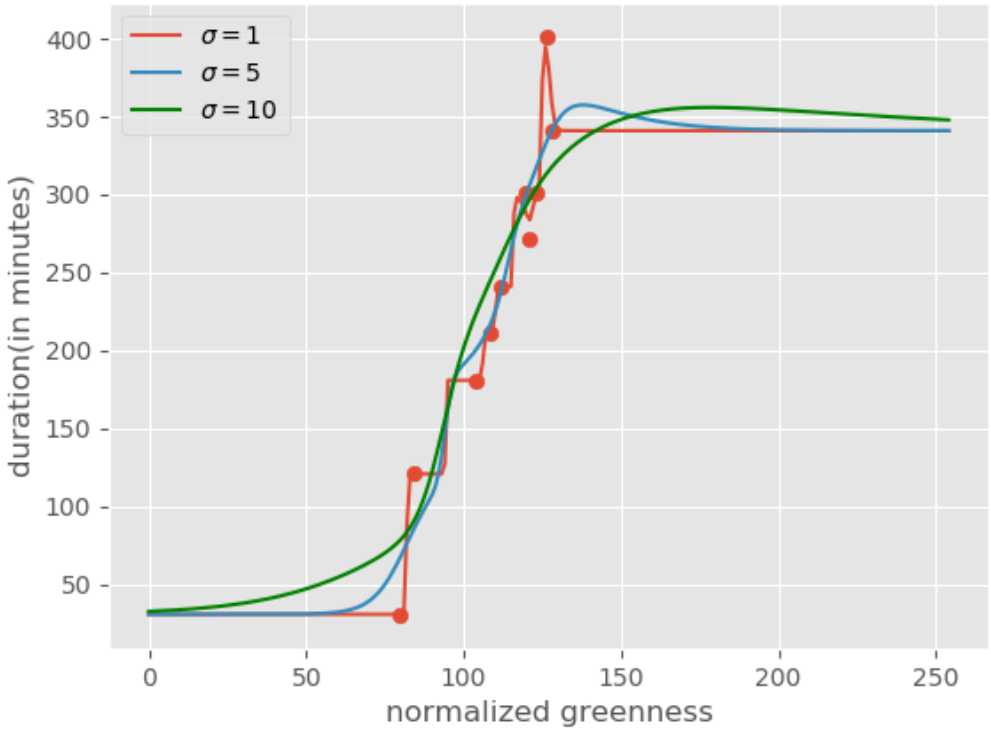

Figure 5: Plot of Kernel Regression

$$\mathcal{E}(h^{(\sigma)}|\mathbb{X})_{\sigma=1} \approx 70.0$$
$$\mathcal{E}(h^{(\sigma)}|\mathbb{X})_{\sigma=5} \approx 850.6$$
$$\mathcal{E}(h^{(\sigma)}|\mathbb{X})_{\sigma=10} \approx 1510.4$$

So choosing $\sigma = 1$ achieves the lowest mean squared error.

Problem 8: Linear Regression using Feature Maps

Answer:
Yes, there is a feature map $\boldsymbol{\phi}$ which allows to approximate the true hypothesis $h^*(\cdot)$ (which satisfies ((9)) by some predictor $h^{(\mathbf{w}_0)}(x) = \mathbf{w}_0^T \boldsymbol{\phi}(x)$ with a suitably chosen weight $\mathbf{w}_0$.

Yes, there is 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 $x \in \mathbb{R}$.

ϕ is a piecewise function.

$$\phi(x) = \begin{cases} 0 & x \notin [0, 10] \\ \phi_{sub}(x) & x \in [0, 10] \end{cases} \tag{19}$$