COMP3670/6670: Introduction to Machine Learning

Release Date. Aug 31st, 2020
Due Date. 23:59pm, Sep 27th, 2020
Total mark 100.

Exercise 1 (15+15+5+5 credits)

Let $\mathbf{A} \in \mathbb{R}^{N\times N}, \Lambda \in \mathbb{R}^{D\times D}$ be symmetric, positive definite. Define $\langle \boldsymbol{x}, \boldsymbol{y}\rangle_{\boldsymbol{A}} := \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{y}$. Define the corresponding norm in the usual fashion.

$$||\boldsymbol{x}||_{\boldsymbol{A}} := \sqrt{\langle x, x\rangle_{\boldsymbol{A}}}$$

We can define $\langle\cdot,\cdot\rangle_\Lambda$ and $||\cdot||_\Lambda$ in the same way. Suppose we are performing linear regression, with a training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$, where for each i, $\boldsymbol{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. We can define the two matrices

$$\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T \in \mathbb{R}^{N\times D}$$

and

$$\boldsymbol{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N.$$

We would like to find $\boldsymbol{\theta} \in \mathbb{R}^D$, such that $\boldsymbol{y} \approx \mathbf{X}\boldsymbol{\theta}$, where the error is measured using $||\cdot||_{\boldsymbol{A}}$. We avoid overfitting by using a weighted regularization term, measured using $||\cdot||_\Lambda$. We define the loss function with regularizer:

$$\mathcal{L}_{\boldsymbol{A},\boldsymbol{\Lambda}}(\theta) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||^2_{\boldsymbol{A}} + ||\boldsymbol{\theta}||^2_\Lambda$$

1. Find the gradient $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{A},\Lambda}(\boldsymbol{\theta})$.

2. Let $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\boldsymbol{A},\Lambda}(\boldsymbol{\theta}) = \mathbf{0}$, and solve for $\boldsymbol{\theta}$.

3. Show that if we choose $\boldsymbol{A} = \boldsymbol{I}$, and $\Lambda = \lambda\boldsymbol{I}$, where $\lambda \in \mathbb{R}$, your answer for 2. agrees with the analytic solution for the standard least squares regression problem with regularization, given by $\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

4. What advantages are there in choosing $||\boldsymbol{\theta}||^2_\Lambda$ over the more standard regularization term $\lambda||\boldsymbol{\theta}||^2_2$?

HINT:

- You may use the property that a symetric positive definite matrix is invertible.

- Since $\boldsymbol{A}$ is symmetric and positive definite, that implies $\langle\cdot,\cdot\rangle_{\boldsymbol{A}}$ is a valid inner product.

Exercise 2 (15+15+15+15 credits)

During linear regression, the goal is to attempt to fit a model to given data. Usually we have a family of models to choose from, and the particular model selected is categorized by parameter θ. We wish to choose the θ that gives us the "best" fit, where "best" is usually defined as the θ that minimises some measure of error. How the error is measured can vastly change the behaviour of the optimal model, along with how easy it is to find an optimal value.

Consider the following linear regression problem in one dimension. We have a collection $\mathcal{D} = \{x_1, \ldots, x_n\}$ of N many points in $\mathbb{R}$, and we want to choose the best representative $\mu \in \mathbb{R}$ for that dataset $\mathcal{D}$. Assume that the points are listed in order,

$$x_1 \leq x_2 \leq \ldots \leq x_n$$

The best choice of μ is one that minimises some loss function $L(\mu, \mathcal{D})$.

1. For the L_2 loss function, $L_2(\mu, D) = \sum_{i=1}^{N} (\mu - x_i)^2$, show that the optimal choice for μ is the mean, $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$.

2. For the L_1 loss function, $L_1(\mu, D) = \sum_{i=1}^{N} |\mu - x_i|$, show that the optimal choice for μ is the median value in the set.
(Hint: Try pairing off opposite terms in L_1, and argue why they must be constant.)

3. For the L_∞ loss function, $L_\infty(\mu, D) = \max_{1 \leq i \leq n} |\mu - x_i|$, show that the optimal choice for μ is $\frac{x_1 + x_n}{2}$, the average of only the largest and smallest data point in the set.
(Hint: Argue why the internal points $x_2, \ldots, x_{n-1}$ are irrelevant.)

4. For the following data sets, compute the optimal representative μ with respect to the L_2, L_1 and L_∞ loss function. Draw your result in a table as shown below. (You don't need to show the calculations for this.) Comment on how sensitive each loss function is to outliers.

$$\mathcal{D}_1 = \{0, 1, 2, 3, 5\}$$
$$\mathcal{D}_2 = \{0, 1, 2, 3, 100\}$$
$$\mathcal{D}_3 = \{0, 10, 30, 35, 100\}$$

μ	L_1	L_2	L_∞
$\mathcal{D}_1$	?	?	?
$\mathcal{D}_2$	?	?	?
$\mathcal{D}_3$	?	?	?

HINT: The middle value in the data set. For odd many points, it is just the middle point. For evenly many points, it is the average of the two middle values.