Assessed Coursework 1

Please submit by 5pm, Friday

Description

- There are 3 questions, worth 20% in total.
- You can use whatever material you can find.
  - However, **do not** copy answers directly from internet.
  - Cite external sources properly.
- You are expected to complete these questions **independently**. Questions of the coursework should be directly addressed to the lecturer.
- You are recommended to use latex to typeset all the answers to the questions.
- Write down the process of your solution. No need to copy the question when answering questions.

Q0: Function Basis Selection

- Scientists have collected a dataset on climate change:
- $\left\{\left(y_i, x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, x_i^{(4)}\right)\right\}$
- y: temperature , $x^{(1)}$: time, $x^{(2)}, x^{(3)}$: longitude, latitude, $x^{(4)}$: CO2 emission.
- **Based on their prior knowledge, they know:**
- 1. temperature changes **periodically** over time
- 2. temperature changes **linearly** with latitude: the higher the latitude, the lower the temperature.
- 3. temperature does **not** change with longitude.
- 4. We do not know how temperature changes with CO2 emission.

Q0: Function Basis Selection

- Therefore, scientists discard $x^{(2)}$ and created a predictive model: $f(\boldsymbol{x}, \boldsymbol{w}) \coloneqq \sum_{i=1,3,4} w_i \phi^{(i)}(x^{(i)})$

- Q0.1 (**2pt**) What would be ideal choices for $\phi^{(1)}, \phi^{(3)}$, $\phi^{(4)}$?
- Choose from linear, polynomial, trigonometric, RBF.
- Explain your choices in a few sentences.

Q0: Function Basis Selection

- Q3.2 (**2pt**) What would **not** happen if scientists accidentally included $x^{(2)}$ in their model $f(\boldsymbol{x}, \boldsymbol{w})$?

Pick one:

- A. Their model overfits
- B. Their model underfits
- C. Training error decreases
- E. Testing error increases

Q1, Gaussian Process

- Gaussian identities we have learned can be used to derive a useful technique called **Gaussian Process**.

- Given a dataset $D \coloneqq \{(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$
- Let us define *a likelihood function* as a n-dim. MVN.
- $p(y_1 \dots y_n | f_1(\boldsymbol{x}_1) \dots f_n(\boldsymbol{x}_n), \sigma) = N_{y_1 \dots y_n}[f_1(\boldsymbol{x}_1) \dots f_n(\boldsymbol{x}_n), \sigma^2 \boldsymbol{I}]$
- where f_i is a function: $R^d \rightarrow R$, $\boldsymbol{I} \in R^{n \times n}$ is an identity matrix.

- Define *a prior* over the functions $f_1 \dots f_n$ as an n-dim. MVN.
- $p(f_1(\boldsymbol{x}_1) \dots f_n(\boldsymbol{x}_n) | \boldsymbol{K}) = N_{f_1 \dots f_n}(\boldsymbol{0}, \boldsymbol{K})$, where $\boldsymbol{K} \in R^{n \times n}$ is a covariance matrix.

Q1, Gaussian Process

- **Q1.1 (2pts)** Write down the expression of $p(\boldsymbol{y}|\sigma, \boldsymbol{K})$
- **Q1.2 (4pts)** Write down the expression of $p(y_1|y_2 \ldots y_n, \boldsymbol{K}, \sigma)$
- **Q1.3 (2pts)** Construct a kernel regression using $\{(y_i, \boldsymbol{x}_i)\}_{i=2}^{n}$ as the training data and use $\boldsymbol{K}$ to construct your kernel function. Predict output value at $\boldsymbol{x}_1$. **Comment on your findings using a few sentences.**
  - "Using $\boldsymbol{K}$ to construct your kernel function" meaning the kernel function is defined as $k(\boldsymbol{x}_i, \boldsymbol{x}_j) := K^{(i,j)}$, the i, j-the element of $\boldsymbol{K}$.
  - Partition $\boldsymbol{K}$ into submatrices $\begin{bmatrix} K_{11}, K_{12} \\ K_{21}, K_{22} \end{bmatrix}$ and express your answers using these submatrices.
- **Q1.4 (2pts)** Comparing to classic least squares using a linear model, what assumption *on our dataset* is *not* required in Gaussian Process? What advantages do we have by not assuming such an assumption?

Q2. Variance-Bias Decomposition

- Given a dataset $D := \{(y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$, assume $y_i = g(x) + \epsilon$, where ϵ is an additive error.
- The prediction function is $f(\boldsymbol{x}; \boldsymbol{w}) := \langle \boldsymbol{w}, \boldsymbol{\phi}(\boldsymbol{x}) \rangle$. $\boldsymbol{\phi}(\boldsymbol{x})$: $R^{\boldsymbol{d}} \to R^b$ and $n > b$.
- There exists a $\boldsymbol{w}^*$ such that $g(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{w}^*)$
- **Q2.1**, (**2pts**) Show that, $\frac{1}{n}\sum_{i \in D} \mathrm{var}[f(\boldsymbol{x}_i; \boldsymbol{w}_{\mathrm{LS}}) | \boldsymbol{x}_i]$ grows as b increases. $\boldsymbol{w}_{\mathrm{LS}}$ is fitted by classic least squares on D.
- **Q2.2**, (**2pts**) Show **in sample error** of $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS}})$ decrease as n increases.
- **Q2.3**, (**2pts**) Comment in a few sentences, on how changing n and b would affect overfitting-ness of $f(\boldsymbol{x}; \boldsymbol{w}_{\mathrm{LS}})$.