Submission Date: 19th November 2021

1 Bias-Variance trade-off

1. "Bias" and "Variance" are terms which can carry many different meanings. Define what they mean in the context of Bias-Variance trade-off. What else can they refer to?

2. Assume you want to predict from MRI images whether a person has cancer or not. You are given only 50 images as they are very expensive to label. The doctors have heard of a very fancy technique called Neural Networks and want to be cutting edge and you to use them.

   (a) How would you expect the Bias and Variance to be and why?

   (b) Are there ways to improve the results?

3. You are given the following data

x_1	x_2	y
-0.8	2.8	-8.5
0.3	-2.2	12.8
1.5	1.1	3.8

   (a) Fit an ordinary least squared model $\hat{h}_1$ without bias term (just two parameters). What are the parameters? What does it predict?

   (b) Fit an ordinary least squared model $\hat{h}_2$ with bias term. What are the parameters? What does it predict?

   (c) The oracle of machine learning has provided you with the true function: $h^*(x) = 5 + 2x_1 - 4x_2$. Generate some more data which will serve as the test set using X_{test}:

x_1	x_2
-2	2
-4	15

   Do you see underfitting, overfitting or neither of them for your models for a), b)?

4. How does regularization prevent overfitting?

2 L2 Regularization

1. Show mathematically that minimizing the loss function for Ridge Regression with parameters θ is equivalent to maximizing the posterior probability $p(\theta|\mathbf{x}, y) = p(y|\ \theta, \mathbf{x})p(\theta)$ (MAP estimation). Assume that $\theta \in \mathrm{R}^N$, $\mathbf{x} = \{\mathbf{x}_1, ..., \mathbf{x}_M\}$ is a set of input variables and $y = \{y_1, ..., y_M\}$ is a set with their respective targets. Moreover:

- $p(\theta)$ is the prior $\prod_{i=1}^{N} \mathcal{N}(\theta_i|0, \sigma_\theta)$
- $p(y_i|\theta, \mathbf{x}_i)$ is the distribution $\mathcal{N}(y_i|\theta^T \mathbf{x}_i, \sigma)$

2. It is possible to establish the same equivalence between Lasso Regression and MAP estimation. How would the prior for θ be $(p(\theta))$?.

3. Based on the previous points, how can we interpret the L1 and L2 regularization from a Bayesian perspective?