CS 135: Machine Learning **Spring 2022**

Problem Set 2

Lecturer: Prof. Peter Chin *Due: March 1, 2022*

- ◇ Please use the code **86748D** to add yourself to the course on Gradescope where you will submit your homework solutions.
- ◇ Please submit your solutions for the written questions (either type up your answers or scan your handwritten solution) in a single PDF on Gradescope under the assignment named `"PS2"` by 23:59PM on the due date. When submitting, please select the correct pages for each question so that we don't miss any part of your solution.
- ◇ For the programming part, please follow the instructions in the shared colab notebook. Save a copy of the notebook to your own Google Drive and include your code in the copy together with necessary graphs, explanations and analysis. Keep the output and make sure the same results can be reproduced by running the colab notebook you submitted in sequence. Please include all necessary files that are required to run the notebook. If it requires additional packages, please specified clearly the procedure to install them. When submitting, submit a link (after enabling sharing) to your finished colab notebook.
- ◇ Late policy: Every student has 7 late days that can be used during the semester. After a student has used all 7 late days, no credit will be given for assignments submitted late without a documented reason. The written part and programming part are considered as a whole when calculating late penalties.

1. (50 points) Written Problems

(a) (10 points) Bishop 3.3 Consider a data set which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$$

Find an expression for the solution $\mathbf{w}^*$ that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise variance and (ii) replicated data points.

(b) (10 points) Bishop 3.4 Consider a linear model of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

Now suppose that Gaussian noise ϵ_i with zero mean and variance σ^2 is added independently to each of the input variables x_i. By making use of $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_j] = \delta_{ij}\sigma^2$, show that minimizing E_D averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter w_0 is omitted from the regularizer.

(c) (10 points) Bishop 3.11 We have seen that, as the size of a data set increases, the uncertainty associated with the posterior distribution over model parameters decreases. Make use of the matrix identity (Appendix C)

$$(\mathbf{M} + \mathbf{v}\mathbf{v}^T)^{-1} = \mathbf{M}^{-1} - \frac{(\mathbf{M}^{-1}\mathbf{v})(\mathbf{v}^T\mathbf{M}^{-1})}{1 + \mathbf{v}^T\mathbf{M}^{-1}\mathbf{v}}$$

to show that the uncertainty of $\sigma_N^2(\mathbf{x})$ associated with the linear regression function given by (3.59) satisfies

$$\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x})$$

(d) (20 points) Consider a Bayesian regression problem with N data points, in the dataset $(\mathbf{X}, \mathbf{t})$, with $\mathbf{t} \in \mathbb{R}^N$ the observed values. Given a data point $\mathbf{x}$ and a parameter values $\mathbf{w} \in \mathbb{R}^D$, our model outputs a prediction $y(\mathbf{x}, \mathbf{w}) \in \mathbb{R}$. Take the likelihood to be a Gaussian with mean $y(\mathbf{x}, \mathbf{w})$ and variance σ_1^2. Take as a prior over weights a multivariate Gaussian with mean 0 and covariance matrix $\sigma_2^2\mathbb{I}$. We can then write

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}) \propto \left(\prod_{n=1}^{N} \mathcal{N}(t_n; y(\mathbf{x}_n, \mathbf{w}), \sigma_1^2) \right) \times \left(\mathcal{N}(\mathbf{w}; 0, \sigma_2^2\mathbb{I}) \right)$$

(i) Write down the explicit formulas for these distributions.

(ii) Finding the value of $\mathbf{w}$ that maximizes this posterior (the MAP point estimate) is equivalent to minimizing the negative log likelihood of the posterior. Take the negative natural log and simplify.

(iii) Discard terms that don't depend on $\mathbf{w}$ and thus show that maximizing the posterior is equivalent to minimizing a sum of squared errors with an l_2 penalty on $\mathbf{w}$.

(iv) Usually we formulate this problem by having a single weight λ on the l_2 penalty. If put the above result into this form, what would λ be?

2. **(50 points) Programming**

   Please follow the instructions in the Google Colab notebooks below. Save a copy to your Google Drive and edit it as instructed. When submitting your homework on Gradescope, please share a link to your finished Colab file. The last edited time on colab is considered your submission time.

   `https://colab.research.google.com/drive/1UOlQaV_nitNchPUTHjAkAz2Is6uZclfi?usp=sharing`

(a) **(20 points) Linear Regression**

(b) **(30 points)** k **Nearest Neighbors**