CS 6140: Machine Learning — Fall 2021— Paul Hand

HW 7
Due: Wednesday November 17, 2021 at 11:59 PM Eastern time via Gradescope.

Names: [Put Your Name(s) Here]

You can submit this homework either by yourself or in a group of 2. You may consult any and all resources. Make sure to justify your answers. If you are working alone, you may either write your responses in LaTeX or you may write them by hand and take a photograph of them. If you are working in a group of 2, you must type your responses in LaTeX. You are encouraged to use Overleaf. Create a new project and replace the tex code with the tex file of this document, which you can find on the course website. To share the document with your partner, click Share > Turn on link sharing, and send the link to your partner. When you upload your solutions to Gradescope, make sure to take each problem with the correct page or image.

Question 1. *Continual learning*

We will consider two machine learning tasks, A and B. Consider two data distributions, $\mathcal{D}_A$ and $\mathcal{D}_B$, over $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. Suppose that for both distributions, the conditional distribution of y given x is the same. Let S_A and S_B be independent samples of size n from $\mathcal{D}_A$ and $\mathcal{D}_B$, respectively. Let $S = S_A \cup S_B$. Consider a Bayesian learning perspective for a parametric statistical model with parameters θ. The prior on the parameters θ is given by the likelihood $p(\theta)$.

(a) In continual learning, one might first train the model on Task A and then continue the learning by training on Task B. MAP estimation of the parameters when training simultaneously on A and B can be performed by using the posterior distribution of training only on A as the prior distribution for training only on B. Show this by establishing that

$$\log p(\theta \mid S) = \log p(S_B \mid \theta) + \log p(\theta \mid S_A) - \log p(S_B).$$

Response:

(b) One approach in practice is to approximate $\log p(\theta \mid S_A)$ as proportional to the expression $-\sum_{j=1}^d F_j(\theta_j - \theta_j^*)^2$, where θ^* is the MAP estimate of θ after training only on task A, and $\{F_j\}$ are positive values that quantify how important each component of θ is. In the context of linear regression, this can give rise to an optimization problem of the following form

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2} \sum_{i=1}^n (y_i - x_i^t \theta)^2 + \lambda \sum_{j=1}^d F_j(\theta_j - \theta_j^*)^2,$$

where the values of (x_i, y_i) come from task B. Write out an analytical expression for the solution to this optimization problem.

Response:

Question 2. *Gradient Descent*

Consider gradient descent with learning rate α applied to the function $f : \mathbb{R}^d \to \mathbb{R}$, $f(\theta) = \frac{1}{2}\theta^t Q\theta$, where $Q \in \mathbb{R}^{d\times d}$ is symmetric.

(a) Let $\theta^{(n)}$ be the value of θ at the nth iteration of gradient descent. Show that

$$\theta^{(n)} = (I - \alpha Q)^n \theta^{(0)}.$$

Response:

(b) Show that if Q is positive definite and $\alpha < \frac{2}{\lambda_{\max}(Q)}$, then $\theta^{(n)}$ converges to zero as $n \to \infty$. Here $\lambda_{\max}(Q)$ is the largest eigenvalue of Q. Hint: directly show that the limit of the expression in (a) is zero. Consider how raising a symmetric matrix to an integer power affects its eigenvalue decomposition.

Response:

(c) Provide an expression of what $\theta^{(n)}$ converges to as $n \to \infty$ if Q is positive semidefinite and if $\alpha < \frac{2}{\lambda_{\max}(Q)}$. Your expression will depend on $\theta^{(0)}$ and aspects of Q.

Response:

Ideas that may help with solving these problems:

Problem 1:

- Bayes Theorem
- MAP Estimation, prior distribution, posterior distribution
- Solution to the Ridge Regression Least Squares
- Diagonal Matrices

Problem 2:

- Eigenvalue Decomposition
- Positive Semidefinite, Positive Definite matrices
- Powers of symmetric matrices
- Relationship of Null Space to Eigenvalue Decomposition
- Matrix Multiplication as the Sum of Outer Products
- **Warm-up for Problem 2 (optional, ungraded):**
  Consider a symmetric matrix $A \in \mathbb{R}^{d\times d}$ and assume that $A = V\Sigma V^T$ where $V \in \mathbb{R}^{d\times d}$ has orthonormal columns and Σ is a diagonal matrix.

  (a) Show that
  $$A^n = V\Sigma^n V^T \qquad \text{for all } n \geq 1.$$

  Response:

  (b) Assume that $\Sigma_{ii} \neq 0$ if $1 \leq i \leq r$ and $\Sigma_{ii} = 0$ if $i > r$. Let $\widehat{\Sigma} \in \mathbb{R}^{r\times r}$ be a diagonal matrix with diagonal elements Σ_{ii} for $1 \leq i \leq r$. Let $V = [V_r, V_n]$ where $V_r \in \mathbb{R}^{d\times r}$ contains the first r columns of V and $V_n \in \mathbb{R}^{d\times(d-r)}$ contains the remaining $d-r$ columns of V. Show that $A = V_r\widehat{\Sigma}V_r^T$.

  Response: