CS 6140: Machine Learning — Fall 2021— Paul Hand

Midterm 2 – revised
Due: Thursday December 2, 2021 at Noon Eastern time via Gradescope.

Name:

There are a total of 110 points available. The exam will be graded out of 100 points.

You must complete this exam by yourself. You may consult any and all resources other than people. Make sure to justify your answers. You may use a computer to perform calculations, such as linear algebra. You may either write your responses in LaTeX or you may write them by hand and take a photograph of them. 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. When you upload your solutions to Gradescope, make sure to take each problem with the correct page or image.

Question 1. *Maximum A Posteriori Estimation*

(20 points) Suppose $y_i \sim \mathcal{N}(\mu, 1)$ for $i = 1 \dots n$. Suppose μ has a Bayesian prior given by a $\mathcal{N}(1, \sigma^2)$ distribution. Find the MAP estimate of μ. Your answer should depend on $\{y_i\}_{i=1\dots n}$ and σ and should be an analytical expression and not an optimization problem.

Response:

Question 2. *Gradient Descent – revised*

(20 points) Consider solving the following minimization problem by gradient descent with step size α. Let $X \in \mathbb{R}^{N \times d}$, $\theta \in \mathbb{R}^d$, and $\lambda > 0$.

$$\min_{\theta} \frac{1}{2}\|X\theta\|^2 + \frac{1}{2}\lambda\|\theta\|^2$$

Show that for any initialization $\theta^{(0)}$, the iterates $\{\theta^{(n)}\}$ given by gradient descent will converge to 0 as $n \to \infty$ if $\alpha < 2/(\lambda + \sigma_{\max}^2(X))$, where $\sigma_{\max}$ is the largest singular value of X. Hint: you may want to use the relationship between the singular values of X and the eigenvalues of X^tX.

Response:

Question 3. *Ridge Regression*

Let $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $\lambda > 0$ and $\theta \in \mathbb{R}^d$. Consider the following optimization problem given by ridge regression:

$$\min_{\theta} \frac{1}{2}\|X\theta - y\|^2 + \frac{1}{2}\lambda\|\theta\|^2.$$

For the following statements, answer whether they are TRUE or FALSE and provide a justification.

(a) (5 points) High values of λ make overfitting more likely.

Response:

(b) (5 points) Least squares linear regression is the case of ridge regression in the limit as $\lambda \to \infty$.

Response:

(c) (5 points) The optimization problem above is convex if $\lambda > 0$.

Response:

(d) (4 points) If $\lambda < 0$, the optimization problem above is necessarily nonconvex.

Response:

Question 4. *Gradient Descent*

Consider gradient descent with step size α on the function $f : \mathbb{R}^d \to \mathbb{R}$. Let $x^{(n)}$ be the nth iterate of gradient descent.

(a) (4 points) Write down the expression relating $x^{(n)}$ to $x^{(n-1)}$.

Response:

(b) (4 points) TRUE or FALSE?
If $f(x) = x^t Q x$ for a matrix $Q \in \mathbb{R}^{d \times d}$, and if α is a small enough positive number, then $x^{(n)}$ will converge to 0 as $n \to \infty$ from any initialization $x^{(0)}$.

Provide a justification. Note: this statement is allowing ANY matrix Q.

Response:

(c) (4 points) TRUE or FALSE? For a general function f, it is always the case that $f(x^{(n+1)}) <= f(x^{(n)})$. If TRUE, provide a justification. If FALSE, present an example where this inequality does not hold and provide a justification.

Response:

(d) (4 points) If gradient descent is not converging for a given α, would it make more sense to try increasing α or decreasing α? Why?

Response:

Question 5. *k Nearest Neighbors*

(a) (5 points) Describe a situation (in the context of classification) where using KNN would be more reasonable than using logistic regression.

Response:

For the following statements, answer whether they are TRUE or FALSE and provide a justification.

(b) (5 points) The bias variance tradeoff does not apply to KNN because it is a nonparametric model for prediction.

Response:

(c) (5 points) Using too large of a value of k for k-nearest neighbors would likely lead to a low bias model.

Response:

Question 6. *Cross Validation*

Consider using k-nearest neighbors to solve a regression problem with the following training data.

x	y
0	2
.1	1.8
.5	1.4
1.2	0.6
2	0.2

(a) (5 points) Consider a regression problem with k-nearest neighbors. If $k = 3$, what value is predicted for the value $x = 1.2$?

Response:

(b) (5 points) If $k = 2$, list all possible values that a k-nearest neighbors predictor could output.

Response:

(c) (10 points) Using leave-one-out cross validation, find the best value of k for KNN regression in this problem. Use square loss to assess validation error.

Response: