Math 189r Fall 2016 Midterm

Name: ________ SOLUTION ________ Time start/end: ________________

Harvey Mudd College's Honor Code is in effect for all students taking this exam. If you feel unclear about any of the following instructions, please ask for clarification.

- This exam should be completed in a contiguous **three hour** period. If you so desire, you may insert a **twenty minute break** at the **one an a half hour mark** to get food, etc., but you are still restricted to the resources available to you during the exam (no internet). You may not write on your exam during the break if you choose to take one.
- No notes, books, computers or calculators will be allowed during the exam or break period except for one sheet of notes (8.5×11", front & back) that you have prepared yourself. When you are finished, staple the note sheet you used to the back of the exam. Please turn in your exam paper into the box outside Professor Gu's office (SHAN3420) no later than Monday, October, 6:30pm.
- Points may be deducted for answers that are not explained clearly.
- Please pace yourself as this exam has four questions, some with multiple parts.

Problem 1	/ 17 points
Problem 2	/ 24 points
Problem 3	/ 34 points
Problem 4	/ 25 points
Total	/ 100 points

Potentially Useful Equations

$$\mathbb{P}_{emp}(x) = \frac{1}{n}\sum_i \delta(x - x_i)$$

$$\int \delta(x-t)f(x)dx = f(t)$$

$$\|\mathbf{x}\|_1 = \sum_i |\mathbf{x}_i|$$

$$\|\mathbf{x}\|_2^2 = \mathbf{x}^\top \mathbf{x}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^\top\right]$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

$$\mathbb{P}(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(x_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_i \mathbf{x}_i$$

$$\hat{\boldsymbol{\Sigma}} = \sum_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$$

$$\mathrm{Cat}(\mathbf{x}|\theta) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

$$\mathbb{P}(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} h(\mathbf{x}) \exp\left(\theta^\top \phi(\mathbf{x})\right)$$

$$\mathbb{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

$$\nabla_{\mathbf{x}}\left(\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x}\right) = (A + A^\top)\mathbf{x} + \mathbf{b}$$

$$\nabla_{\mathbf{x}}^2\left(\mathbf{x}^\top A\mathbf{x} + \mathbf{b}^\top \mathbf{x}\right) = (A + A^\top)$$

1. (17 Points) Consider the Ridge Regression optimization problem

$$\text{minimize}: f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \|\mathbf{\Gamma}\mathbf{x}\|_2^2.$$

(a) Solve for the optimal $\mathbf{x}^\star$ in closed form.

At optimality we know by the Karush-Kuhn-Tucker conditions that $\nabla f(\mathbf{x}^\star) = 0$. Then since $f(\mathbf{x}) = (\mathbf{A}\mathbf{x} - \mathbf{b})^\top(\mathbf{A}\mathbf{x} - \mathbf{b}) + \mathbf{x}^\top\mathbf{\Gamma}^\top\mathbf{\Gamma}\mathbf{x} = \mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{x}^\top\mathbf{A}^\top\mathbf{b} + \mathbf{b}^\top\mathbf{b} + \mathbf{x}^\top\mathbf{\Gamma}^\top\mathbf{\Gamma}\mathbf{x}$ we have

$$\nabla f = 2\mathbf{A}^\top\mathbf{A}\mathbf{x} - 2\mathbf{A}^\top\mathbf{b} + 2\mathbf{\Gamma}^\top\mathbf{\Gamma}\mathbf{x} = 0$$

or

$$(\mathbf{A}^\top\mathbf{A} + \mathbf{\Gamma}^\top\mathbf{\Gamma})\mathbf{x}^\star = \mathbf{A}^\top\mathbf{b}.$$

If $\mathbf{A}^\top\mathbf{A} + \mathbf{\Gamma}^\top\mathbf{\Gamma}$ is invertible then

$$\mathbf{x}^\star = (\mathbf{A}^\top\mathbf{A} + \mathbf{\Gamma}^\top\mathbf{\Gamma})^{-1}\mathbf{A}^\top\mathbf{b}.$$

(b) Consider a dataset $\mathcal{D} = \{(0,1), (1,2), (2,1), (3,2)\}$. Construct a matrix A and $\mathbf{b}$ from this dataset $\mathcal{D}$ and compute the Ridge estimate with $\mathbf{\Gamma} = \mathbf{I}$. Note you may leave your solution in the form of the inverse of a matrix times a vector instead of computing the inverse by hand. All other operations must be simplified.

Accounting for the bias term we have

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

so

$$\mathbf{A}^\top\mathbf{A} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \qquad \mathbf{A}^\top\mathbf{b} = \begin{bmatrix} 6 \\ 10 \end{bmatrix} \qquad \mathbf{\Gamma}^\top\mathbf{\Gamma} = \mathbf{I}^\top\mathbf{I} = \mathbf{I},$$

giving the ridge estimate

$$\mathbf{x}^\star = (\mathbf{A}^\top\mathbf{A} + \mathbf{\Gamma}^\top\mathbf{\Gamma})^{-1}\mathbf{A}^\top\mathbf{b} = \begin{bmatrix} 5 & 6 \\ 6 & 15 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 10 \end{bmatrix}.$$

2. (24 Points - 3 each) Indicate whether each of the following statements is true (T) or not necessarily true (X). Be careful with your answers as no partial credit will be awarded on this problem.

(a) $\boxed{\text{X}}$ The following is true: $\text{cov}[A\mathbf{x} + \mathbf{b}] = A^\top \text{cov}[\mathbf{x}]A$.

(b) $\boxed{\text{T}}$ Consider minimizing $-\ell(\mathbf{w}, \mathcal{D}_{train}) + \lambda\|\mathbf{w}\|_2^2$ where $\ell(\mathbf{w}, \mathcal{D}) = \frac{1}{n}\sum_i y_i \log \sigma(\mathbf{w}^\top \mathbf{x}_i) + (1 - y_i)\log(1 - \sigma(\mathbf{w}^\top \mathbf{x}_i))$ is the average log-likelihood on a dataset $\mathcal{D}$ for the ℓ_2-regularized logistic regression model. If the training data is linearly separable, might some weights w_j become infinite if $\lambda = 0$? (Murphy 8.6.c)

(c) $\boxed{\text{X}}$ Adding a zero-mean Gaussian Prior on each of the weights in a Logistic Regression model will encourage sparser weights than a zero-mean Laplace prior in a Maximum-a-Posteriori estimation of the optimal parameters.

(d) $\boxed{\text{T}}$ Minimizing $\|X\mathbf{w} - \mathbf{y}\|_2^2$ and maximizing the likelihood $\prod_i \mathcal{N}(\mathbf{y}_i|\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$ all result in the Normal Equations, the optimal solution $X^\top X\mathbf{w}^\star = X^\top \mathbf{y}$, in the context of linear regression.

(e) $\boxed{\text{X}}$ A covariance matrix $\mathbf{\Sigma}$ could potentially have a negative eigenvalue.

(f) $\boxed{\text{T}}$ The level sets of a multivariate Gaussian density are always ellipses of the form $\{\mathbf{x} : \mathbf{x}^\top A\mathbf{x} = k\}$.

(g) $\boxed{\text{X}}$ The Multivariate Normal Distribution is not in the exponential family.

(h) $\boxed{\text{T}}$ The goal of the Support Vector Machine is to maximize the margin, defined as the distance of the closest examples from the decision boundary.

3. (34 Points) Consider a Poisson distributed $X \sim \text{Poi}(\lambda)$ defined over $X \in \{0, 1, 2, \dots\}$ with probability mass function

$$\text{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

Note that $\mathbb{E}[X] = \lambda$.

(a) Show that the Poisson distribution is in the Exponential Family.

We have

$$\begin{aligned}\text{Poi}(x|\lambda) &= e^{-\lambda}\frac{\lambda^x}{x!}\\ &= \frac{e^{-\lambda}}{x!}\exp\left(x \log \lambda\right),\end{aligned}$$

so the Poisson distribution is in the exponential family with

$$Z(\theta) = e^{\lambda} \qquad h(x) = \frac{1}{x!} \qquad \phi(x) = x \qquad \theta = \log \lambda.$$

(b) Consider creating a Generalized Linear Model from the Poisson distribution to model some count data. This will work since we showed that the distribution is an exponential family distribution. What is the distribution of the predicted value $\hat{y}$ given a datapoint $\mathbf{x}_i$ and weights $\mathbf{w}$?

For this linear model we assign $\theta = \mathbf{w}^\top \mathbf{x} = \log \lambda$ so $\lambda = \exp\left(\mathbf{w}^\top \mathbf{x}\right)$ or

$$\begin{aligned}\mathbb{P}(y|\mathbf{x}, \mathbf{w}) &= \text{Poi}\left(y|e^{\mathbf{w}^\top \mathbf{x}}\right)\\ &= \frac{e^{-\exp(\mathbf{w}^\top \mathbf{x})}}{y!}\exp\left(y \cdot \mathbf{w}^\top \mathbf{x}\right).\end{aligned}$$

(c) Derive an expression for the log-likelihood of a dataset $\mathcal{D}$, $\log \mathbb{P}(\mathcal{D}|\mathbf{w})$. Assume the data is identically and independently distributed.

Assuming our data is identically and independently distributed as we always do for generalized linear models we have by part (b)

$$\begin{aligned}
\log \mathbb{P}(\mathcal{D}|\mathbf{w}) &= \log \prod_i \mathbb{P}(y_i|\mathbf{x}_i, \mathbf{w}) \\
&= \sum_i \log \mathbb{P}(y_i|\mathbf{x}_i, \mathbf{w}) \\
&= \sum_i \log \left(\frac{e^{-\exp(\mathbf{w}^\top \mathbf{x}_i)}}{y_i!} \exp\left(y_i \cdot \mathbf{w}^\top \mathbf{x}_i\right) \right) \\
&= \sum_i y_i \cdot \mathbf{w}^\top \mathbf{x}_i - e^{\mathbf{w}^\top \mathbf{x}_i} - \log y_i!.
\end{aligned}$$

(d) Suppose we place an isotropic ($\mathbf{\Sigma} = \sigma^2 \mathbf{I}$) Gaussian prior on the weights $\mathbf{w}$. Derive an expression which, when maximized, would maximize $\log \mathbb{P}(\mathbf{w}|\mathcal{D})$ (basically ignore constants). *Hint:* think about problem 1.a from homework 1. You may use results from that part of the assignment. You may introduce an auxillary variable λ which somehow relates to σ^2 (and you don't need to define that relationship exactly).

Adding the prior corresponds to adding $\lambda \mathbf{w}^\top \mathbf{w}$ to the likelihood found in part (c). Dropping the constant term $\sum_i \log y_i!$ and flipping to a minimization problem we have that maximizing $\log \mathbb{P}(\mathbf{w}|\mathcal{D})$ is equivalent to solving

$$\text{minimize: } f(\mathbf{w}) = \sum_i e^{\mathbf{w}^\top \mathbf{x}_i} - y_i \cdot \mathbf{w}^\top \mathbf{x}_i + \lambda \|\mathbf{w}\|_2^2$$

(e) Suppose we want to compute a maximum-a-posteriori estimate of $\mathbf{w}$ given the prior from the previous problem. State the optimization problem we are trying to solve. Is this solvable in closed form? If so, solve it. Otherwise, compute the gradient of the objective function with respect to $\mathbf{w}$.

This problem is not solvable in closed form (see the gradient expression below!) but it is a convex problem. Taking the gradient, then, we have

$$\begin{aligned}
\nabla f(\mathbf{w}) &= \sum_i x_i e^{\mathbf{w}^\top \mathbf{x}_i} - y_i x_i + 2\lambda \mathbf{w} \\
&= \sum_i \left(e^{\mathbf{w}^\top \mathbf{x}_i} - y_i \right) x_i + 2\lambda \mathbf{w}.
\end{aligned}$$

4. (25 Points) Suppose we want to model where we have some input data $\mathbf{X}$, with each datapoint corresponding to some observed function output $\mathbf{f}$. We also have points $\mathbf{X}_\star$ with which we want to predict what the function output $\mathbf{f}_\star$ will be. Our modelling assumption is this:

   (1) The function outputs are jointly normal with mean $\boldsymbol{\mu}$ and covariance $\mathbf{K}$ such that

$$\begin{pmatrix}\mathbf{f}\\ \mathbf{f}_\star\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}_\star\end{pmatrix}, \begin{pmatrix}\mathbf{K} & \mathbf{K}_\star\\ \mathbf{K}_\star^\top & \mathbf{K}_{\star\star}\end{pmatrix}\right)$$

   (2) The mean $\boldsymbol{\mu} = \boldsymbol{\mu}(X) = \big(m(\mathbf{x}_i), \ldots, m(\mathbf{x}_n)\big)$ for $m : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X}$ is the space your data live in (points on the line, locations on earth, molecules, etc.). Notationally, $\boldsymbol{\mu}_\star = \boldsymbol{\mu}(X_\star)$.

   (3) The covariance between two datapoints $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Later in the course we'll call this $\kappa(\cdot, \cdot)$ a valid kernel. This implies that $\mathbf{K} \in \mathbb{R}^{n \times n}$, $\mathbf{K}_\star \in \mathbb{R}^{n \times n_\star}$, and $\mathbf{K}_{\star\star} \in \mathbb{R}^{n_\star \times n_\star}$.

   (a) Compute the posterior predictive distribution $p(\mathbf{f}_\star | \mathbf{X}_\star, \mathbf{X}, \mathbf{f})$ when we assume that the observed $\mathbf{f}$ has no noise. *Hint:* consider conditioning a Gaussian.

   Using the Gaussian Conditioning Theorem we have the conditional distribution

$$\begin{aligned}\mathbb{P}(\mathbf{f}_\star | \mathbf{X}_\star, \mathbf{X}, \mathbf{f}) &= \mathcal{N}(\mathbf{f}_\star | \boldsymbol{\mu}_\star, \boldsymbol{\Sigma}_\star), \text{ where}\\ \boldsymbol{\mu}_\star &= \boldsymbol{\mu}(\mathbf{X}_\star) + \mathbf{K}_\star^\top \mathbf{K}^{-1}(\mathbf{f} - \boldsymbol{\mu}(\mathbf{X})) \text{ and}\\ \boldsymbol{\Sigma}_\star &= \mathbf{K}_{\star\star} - \mathbf{K}_\star^\top \mathbf{K}^{-1} \mathbf{K}_\star\end{aligned}$$

   (b) Compute the posterior predictive distribution $p(\mathbf{f}_\star | \mathbf{X}_\star, \mathbf{X}, \mathbf{y})$ where we assume our observations $\mathbf{y} = f(\mathbf{x}) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. *Hint:* this assumption will only change $\mathbf{f}$ to $\mathbf{y}$ and $\mathbf{K}$ (not $\mathbf{K}_\star$ or $\mathbf{K}_{\star\star}$) into $\mathbf{K} + \sigma^2\mathbf{I}$ from the equation in (1). For this part assume the mean $\boldsymbol{\mu} = \mathbf{0}$.

   From (a) we have (adding the noise to the $\mathbf{K}$ term and setting $\boldsymbol{\mu} = \mathbf{0}$)

$$\begin{aligned}\mathbb{P}(\mathbf{f}_\star | \mathbf{X}_\star, \mathbf{X}, \mathbf{y}) &= \mathcal{N}(\mathbf{f}_\star | \boldsymbol{\mu}_\star, \boldsymbol{\Sigma}_\star), \text{ where}\\ \boldsymbol{\mu}_\star &= \mathbf{K}_\star^\top \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{y} \text{ and}\\ \boldsymbol{\Sigma}_\star &= \mathbf{K}_{\star\star} - \mathbf{K}_\star^\top \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{K}_\star\end{aligned}$$

   (c) Show that when predicting at only one location $\mathbf{x}_\star$ the mean of the predictive distribution from the previous part $\overline{\mathbf{f}} = \sum_{i=1}^n \alpha_i \kappa(\mathbf{x}_i, \mathbf{x}_\star)$ with $\alpha = (\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$ being a constant vector depending on your training data and kernel. Armed with this knowledge, how might be we restrict the predicted function $\overline{\mathbf{f}}(\mathbf{x})$ to be periodic? *Hint:* note that $\kappa(\mathbf{x}_i, \mathbf{x}_\star)$ is function of $\mathbf{x}_\star$ and the sum of periodic functions is periodic.

   From (b) we have for one test input $\mathbf{x}_\star$ the mean $\overline{\mathbf{f}} = \boldsymbol{\mu} = \mathbf{k}_\star^\top \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{y}$. If we let $\boldsymbol{\alpha} = \left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1} \mathbf{y}$ we have $\overline{\mathbf{f}} = \sum_{i=1}^n \mathbf{k}_{\star i} \boldsymbol{\alpha}_i = \sum_{i=1}^n \boldsymbol{\alpha}_i \kappa(\mathbf{x}_i, \mathbf{x}_\star)$ from assumption (3). Since $\kappa(\mathbf{x}_i, \cdot)$ is just a function of $\mathbf{x}_\star$ we can just restrict $\kappa(\cdot, \cdot)$ to be periodic in either one of it's arguments. Since $\overline{\mathbf{f}}$ is a linear combination of periodic functions under this assumption we have that $\overline{\mathbf{f}}$ is periodic in $\mathbf{x}_\star$ as desired.