Notes:

1. The midterm will have about 10 marks MCQ, and about 20 marks description questions will be provided on the actual midterm for you to write your answers.

2. The midterm is meant to be educational, and as such some questions could be quite challenging. Use your time wisely to answer as much as you can.

1. **[13 points] Generalized Linear Models**

   Recall that generalized linear models assume that the response variable y (conditioned on x) is distributed according to a member of the exponential family:

   $$P(y;\eta) = b(y)\exp(\eta T(y) - a(\eta)),$$

   where $\eta = \theta^T x$. For this problem, we will assume $\eta \in \mathbb{R}$.

   (a) **[10 points]** Given a training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$, the loglikelihood is given by

   $$\ell(\theta) = \sum_{i=1}^m \log p(y^{(i)} \mid x^{(i)}; \theta).$$

   Give a set conditions on $b(y)$, $T(y)$, and $a(\eta)$ which ensure that the loglikelihood is a concave function of θ (and thus has a unique maximum). Your conditions must be reasonable, and should be as weak as possible. (E.g., the answer "any $b(y)$, $T(y)$, and $a(\eta)$ so that $\ell(\theta)$ is concave" is not reasonable. Similarly, overly narrow conditions, including ones that apply only to specific GLIMs, are also not reasonable.)

   (b) **[3 points]** When the response variable is distributed according to a Normal distribution (with unit variance), we have $b(y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}}$, $T(y) = y$, and $a(\eta) = \frac{\eta^2}{2}$. Verify that the condition(s) you gave in part (a) hold for this setting.

2. **[15 points] Bayesian linear regression**

   Consider Bayesian linear regression using a Gaussian prior on the parameters $\theta \in \mathbb{R}^{n+1}$. Thus, in our prior, $\theta \sim \mathcal{N}(\vec{0}, \tau^2 I_n)$, where $\tau^2 \in \mathbb{R}$, and I_{n+1} is the $n+1$-by-$n+1$ identity matrix. Also let the conditional distribution of $y^{(i)}$ given $x^{(i)}$ and θ be $\mathcal{N}(\theta^T x^{(i)}, \sigma^2)$, as in our usual linear least-squares model.[1] Let a set of m IID training examples be given (with $x^{(i)} \in \mathbb{R}^{n+1}$). Recall that the MAP estimate of the parameters θ is given by:

   $$\theta_{MAP} = \arg\max_\theta \left(\prod_{i=1}^m p(y^{(i)} | x^{(i)}, \theta) \right) p(\theta)$$

   Find, in closed form, the MAP estimate of the parameters θ. For this problem, you should treat τ^2 and σ^2 as fixed, known, constants. [Hint: Your solution should involve deriving something that looks a bit like the Normal equations.]

3. **[18 points] Kernels**

In this problem, you will prove that certain functions K give valid kernels. Be careful to justify every step in your proofs. Specifically, if you use a result proved either in the lecture notes or homeworks, be careful to state exactly which result you're using.

(a) **[8 points]** Let $K(x, z)$ be a valid (Mercer) kernel over $\mathbb{R}^n \times \mathbb{R}^n$. Consider the function given by

$$K_e(x, z) = \exp(K(x, z)).$$

Show that K_e is a valid kernel. [Hint: There are many ways of proving this result, but you might find the following two facts useful: (i) The Taylor expansion of e^x is given by $e^x = \sum_{j=0}^{\infty} \frac{1}{j!} x^j$ (ii) If a sequence of non-negative numbers $a_i \geq 0$ has a limit $a = \lim_{i \to \infty} a_i$, then $a \geq 0$.]

(b) **[8 points]** The Gaussian kernel is given by the function

$$K(x, z) = e^{-\frac{||x-z||^2}{\sigma^2}},$$

where $\sigma^2 > 0$ is some fixed, positive constant. We said in class that this is a valid kernel, but did not prove it. Prove that the Gaussian kernel is indeed a valid kernel. [Hint: The following fact may be useful. $||x - z||^2 = ||x||^2 - 2x^T z + ||z||^2$.]

4. **[18 points] One-class SVM**

Given an unlabeled set of examples $\{x^{(1)}, \ldots, x^{(m)}\}$ the *one-class SVM algorithm* tries to find a direction w that maximally separates the data from the origin. [2]

More precisely, it solves the (primal) optimization problem:

$$\begin{aligned} \min_w \quad & \frac{1}{2} w^\top w \\ \text{s.t.} \quad & w^\top x^{(i)} \geq 1 \qquad \text{for all } i = 1, \ldots, m \end{aligned}$$

A new test example x is labeled 1 if $w^\top x \geq 1$, and 0 otherwise.

(a) **[9 points]** The primal optimization problem for the one-class SVM was given above. Write down the corresponding dual optimization problem. Simplify your answer as much as possible. In particular, w should not appear in your answer.

(b) **[4 points]** Can the one-class SVM be kernelized (both in training and testing)? Justify your answer.

(c) **[5 points]** Give an SMO-like algorithm to optimize the dual. I.e., give an algorithm that in every optimization step optimizes over the smallest possible subset of variables. Also give in closed-form the update equation for this subset of variables. You should also justify why it is sufficient to consider this many variables at a time in each step.

[5 points] Suppose an ℓ_1-regularized SVM (with regularization parameter $C > 0$) is trained on a dataset that is linearly separable. Because the data is linearly separable, to minimize the primal objective, the SVM algorithm will set all the slack variables to zero. Thus, the weight vector w obtained will be the same no matter what regularization parameter C is used (so long as it is strictly bigger than zero). True or false?

[5 points] Consider using hold-out cross validation (using 70% of the data for training, 30% for hold-out CV) to select the bandwidth parameter τ for locally weighted linear regression. As the number of training examples m increases, would you expect the value of τ selected by the algorithm to generally become larger, smaller, or neither of the above? For this problem, assume that (the expected value of) y is a non-linear function of x.

[5 points] Consider a feature selection problem in which the mutual information $MI(x_i, y) = 0$ for all features x_i. Also for every subset of features $S_i = \{x_{i_1}, \cdots, x_{i_k}\}$ of size $< n/2$ we have $MI(S_i, y) = 0$.[3] However there is a subset S^* of size exactly $n/2$ such that $MI(S^*, y) = 1$. I.e. this subset of features allows us to predict y correctly. Of the three feature selection algorithms listed below, which one do you expect to work best on this dataset?

i. Forward Search.

ii. Backward Search.

iii. Filtering using mutual information $MI(x_i, y)$.

iv. All three are expected to perform reasonably well.