DS-GA 1003 : Machine Learning

Spring 2020: Midterm Practice Exam (100 Minutes)

Answer the questions in the spaces provided. If you run out of room for an answer, use the blank page at the end of the test. Please **don't miss the last question**, on the back of the last test page.

Name: ______________________________

NYU NetID: ______________________________

Question	Points	Score
Variants of SVM objective	2	
Hypothesis Spaces	4	
L1/L2 Regularization	1	
Stopping Rules	5	
(S)GD	3	
Sparsity in Lasso and SVM	3	
Determining Kernels for Classification	9	
Kernelized Regression	8	
Total:	35	

1. (2 points) Suppose you have a function

$$\varphi(c) = \arg\min_{w \in \mathbf{R}^d} \frac{c}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i[w^T x_i]\right) + \frac{1}{2}||w||^2.$$

where $c > 0$. Show how we could use this $\varphi(c)$ to find a minimizer $w^* \in \mathbf{R}^d$ of the following objective function (where $\lambda > 0$):

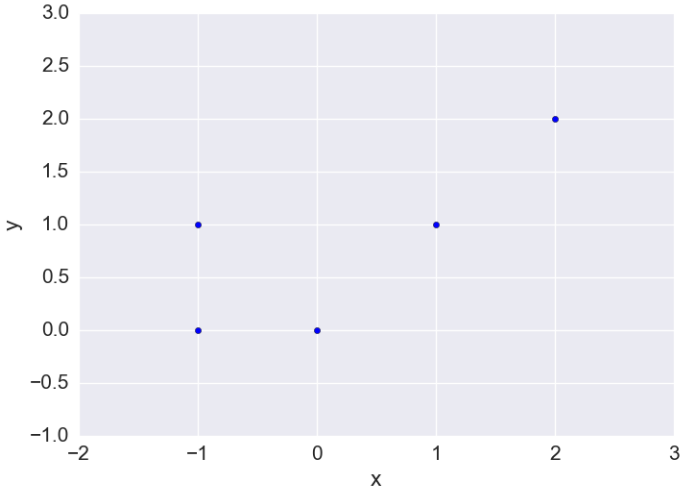
$$J(w) = \frac{1}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i[w^T x_i]\right) + \lambda||w||^2$$

Solution: Note

$$\begin{aligned}
&\frac{1}{2\lambda}\left(\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i[w^T x_i]\right) + \lambda||w||^2\right) \\
&\quad = \frac{1/(2\lambda)}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i[w^T x_i]\right) + \frac{\lambda}{2\lambda}||w||^2 \\
&\quad = \frac{1/(2\lambda)}{n} \sum_{i=1}^{n} \max\left(0, 1 - y_i[w^T x_i]\right) + \frac{1}{2}||w||^2
\end{aligned}$$

Thus, let $c = \frac{1}{2\lambda}$. Then $\varphi(\frac{1}{2\lambda})$ will return a minimizer of objective $J_2(w)$.

2. Let $\mathcal{X} = \mathcal{Y} = \mathcal{A} = \mathbb{R}$. Suppose you receive the (x, y) data points (-1,1), (-1,0), (0,0), (1,1), and (2,2).



(a) Assume we're using the 0-1 loss function $\ell(a, y) = \mathbf{1}(a \neq y)$.

i. (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_c$ of constant functions. Give an empirical risk minimizer $\hat{f}(x)$.

Solution: $\hat{f}(x) = 0$ or 1.

ii. (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_c$ of constant functions. What is $\hat{R}(\hat{f})$, the empirical risk of $\hat{f}$, where $\hat{f}$ is an empirical risk minimizer.

Solution: 3/5.

iii. (1 point) Suppose we restrict to the hypothesis space $\mathcal{F}_\ell$ of linear functions. Give an empirical risk minimizer $\hat{f}(x)$.

Solution: $\hat{f}(x) = x$.

(b) Now assume we're using the absolute loss function $\ell(a, y) = |a - y|$.

i. (1 point) What is the minimum empirical risk achievable over the hypothesis space of all functions?

Solution: 1/5.

3. (a) (1 point) Consider the following version of the elastic-net objective:

$$J(w) = \frac{1}{n}\|Xw - y\|_2^2 + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2.$$

Our training data is $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, sampled i.i.d. from some distribution P. As usual, the design matrix $X \in \mathbb{R}^{n\times d}$ has x_i as its ith row, and $y \in \mathbb{R}^n$ has y_i as its i'th coordinate. Which ONE of the following hyperparameter settings is most likely to give a sparse solution?

□ $\lambda_1 = 0, \lambda_2 = 1$ ■ $\lambda_1 = 1, \lambda_2 = 0$ □ $\lambda_1 = 0, \lambda_2 = 0$

4. The penalized empirical risk for $f \in \mathcal{F}$ and dataset $\mathcal{D}$ is given by

$$J(f; \mathcal{D}) = \hat{R}(f; \mathcal{D}) + \lambda \Omega(f),$$

where $\Omega : \mathcal{F} \to [0, \infty)$ is a regularization function, $\lambda > 0$ is a regularization parameter, and

$$\hat{R}(f; \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} \ell(f(x), y)$$

is the empirical risk of f for the data $\mathcal{D}$, where $|\mathcal{D}|$ is the size of the set $\mathcal{D}$.

(a) (1 point) Suppose we use an iterative descent method to minimize $J(f; \mathcal{D})$ for some training data $\mathcal{D}$. Let $f^{(i)}$ be the prediction function at the i'th iteration. If our goal is to find a minimizer $\hat{f} \in \arg\min_{f \in \mathcal{F}} J(f; \mathcal{D})$, which ONE of the following is the better stopping condition?

□ $\hat{R}(f^{(i)}; \mathcal{D}) - \hat{R}(f^{(i+1)}; \mathcal{D}) < \epsilon$, for some appropriately chosen $\epsilon > 0$.

■ $J(f^{(i)}; \mathcal{D}) - J(f^{(i+1)}; \mathcal{D}) < \epsilon$**, for some appropriately chosen** $\epsilon > 0$**.**

(b) (1 point) A friend reminds us that our real goal is to find an f that has small risk, i.e. a small value of $R(f) = \mathbb{E}\ell(f(x), y)$. Suppose we have found $\tilde{f}$ using $\mathcal{D}$ and we have an independent validation set $\mathcal{D}_{\text{val}}$ from the same distribution as $\mathcal{D}$. Select ALL of the following that are unbiased estimators of $R(\tilde{f})$:

□ $J(\tilde{f}, \mathcal{D}_{val})$ □ $\hat{R}(\tilde{f}, \mathcal{D})$ □ $J(\tilde{f}, \mathcal{D})$ ■ $\hat{R}(\tilde{f}, \mathcal{D}_{val})$

(c) (2 points) A friend thinks that we are running too many iterations of the optimization procedure to minimize $J(f; \mathcal{D})$, when our real goal is to find f with small risk. Let

$$f^* \in \arg\min_f \mathbb{E}\ell(f(x), y)$$

$$f_{\mathcal{F}} \in \arg\min_{f \in \mathcal{F}} \mathbb{E}\ell(f(x), y)$$

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} J(f; \mathcal{D})$$

and again let $f^{(i)}$ be the prediction function at the current iteration. Select ALL of the following that would support your friend's claim that it's time to stop the optimization algorithm:

■ $J(f^{(i)}, \mathcal{D}) = J(\hat{f}, \mathcal{D})$

■ $R(f^{(i)}) < R(\hat{f})$

□ $\hat{R}(f^{(i)}, \mathcal{D}) < \hat{R}(\hat{f}, \mathcal{D})$

□ $R(\hat{f}) - R(f_{\mathcal{F}})$ is significantly smaller than $R(f_{\mathcal{F}}) - R(f^*)$

(d) (1 point) Your friend suggests using your validation data $\mathcal{D}_{val}$ for "early stopping." Which ONE of the following is the BEST suggestion for an early stopping rule that **you could use in practice**:

- ■ $\hat{R}(f^{(i-100)}, \mathcal{D}_{val}) - \hat{R}(f^{(i)}, \mathcal{D}_{val}) < \epsilon$ **for some appropriately chosen** $\epsilon > 0$**.**
- □ $J(f^{(i-100)}, \mathcal{D}_{val}) - J(f^{(i)}, \mathcal{D}_{val}) < \epsilon$ for some appropriately chosen $\epsilon > 0$.
- □ $\hat{R}(f^{(i)}, \mathcal{D}_{val}) - \hat{R}(\hat{f}, \mathcal{D}_{val}) < \epsilon$ for some appropriately chosen $\epsilon > 0$.
- □ $J(f^{(i)}, \mathcal{D}_{val}) - J(\hat{f}, \mathcal{D}_{val}) < \epsilon$ for some appropriately chosen $\epsilon > 0$.

5. Decide whether the following statements apply to full batch gradient descent (GD), minibatch GD, neither, or both. Assume we're minimizing a differentiable, convex objective function $J(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, and we are currently at w_t, which is not a minimum. For full batch GD, take $v = \nabla_w J(w_t)$, and for minibatch GD take v to be a minibatch estimate of $\nabla_w J(w_t)$ based on a random sample of the training data.

(a) (1 point) For any step size $\eta > 0$, after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$, we must have $J(w_{t+1}) < J(w_t)$. (Choose ONE answer below.)

□ Full batch □ Minibatch □ Both ■ **Neither**

(b) (1 point) There must exist some $\eta > 0$ such that after applying the update rule $w_{t+1} \leftarrow w_t - \eta v$ we have $J(w_{t+1}) < J(w_t)$. (Choose ONE answer below.)

■ **Full batch** □ Minibatch □ Both □ Neither

(c) (1 point) v is an unbiased estimator of the full batch gradient. (Choose ONE answer below.)

□ Full batch □ Minibatch ■ **Both** □ Neither

6. We have discussed two methods that we claim can give sparsity: Lasso and SVM. In this question you'll compare the "sparsity" achieved through these methods.

   (a) (1 point) Suppose $f : \mathbf{R}^d \to \mathbf{R}$ is a prediction function attained by linear least squares Lasso regression. Give an expression for $f(x)$ in terms of a vector that may be sparse in the lasso context, and specify which vector may be sparse.

   Solution: $f(x) = w^T x$, where w may be sparse.

   (b) (2 points) Suppose $f : \mathbf{R}^d \to \mathbf{R}$ is a prediction function (i.e. score function) from a linear SVM. Give an expression for $f(x)$ in terms of a vector that may be sparse in the SVM context, and specify which vector may be sparse.

   Solution: $f(x) = \sum_{i=1}^{n} \alpha_i x_i^T x$, where $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ may be sparse.

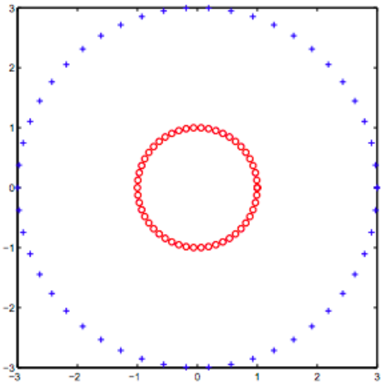
7. (2 points) (a) (1 point) Show that the following kernel function is a Mercer kernel (i.e. it represents an inner product):

$$k(x, y) = \frac{x^T y}{\|x\|\|y\|},$$

where $x, y \in \mathbf{R}^d$.

Solution: Take $\varphi(x) = \frac{x}{||x||}$

(b) (2 points) Consider the binary classification problem shown in Figure b: Denote the input space by $\mathcal{X} = \{(x_1, x_2) \in \mathbf{R}^2\}$. Give a feature mapping for which a linear classifier could perfectly separate the two classes shown.



Solution: Take $(x, y) \mapsto (1, x, y, x^2, xy, y^2)$

(c) (2 points) For the classification problem in Figure b, check all classifiers that could perfectly separate the classes:

□ linear SVM

□ SVM with quadratic kernel

■ **SVM with radial basis functions**

(d) (2 points) Suppose we fit a hard-margin SVM to N data points, and we have 2 data points "on the margin". If we add a new data point to the training set and refit the SVM, what's the largest number of data points that could end up "on the margin". Support your answer (a picture could suffice).

Solution:

8. Consider the regression setting in which $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{N}$ and $\mathcal{A} = \mathbb{R}$ with a linear hypothesis space $\mathcal{F} = \{f(x) = w^T x \mid w \in \mathbb{R}^d\}$ and the loss function

$$\ell(\hat{y}, y) = -y\hat{y} + \exp(\hat{y}) + \log(y!),$$

where $\hat{y}$ is the action and y is the outcome. Consider the objective function

$$J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w^T x_i, y_i) + \lambda ||w||^2$$

(a) (4 points) Provide a kernelized objective function $J_k(\alpha) : \mathbf{R}^n \to \mathbf{R}$. You may write your answer in terms of the Gram matrix $K \in \mathbf{R}^{n\times n}$, defined as $K_{ij} = x_i^T x_j$. (Hint: Recall the representer theorem).

Solution: By the Representer theorem (you should be able to justify its application, but you do not need to for full credit), we know that for $w^* = \sum_{i=1}^n \alpha_i^* x_i = X^T \alpha^*$, where $X \in \mathbf{R}^{n\times d}$ is the usual design matrix. So $w^{*T} x_i = (K\alpha^*)_i$, where $K = XX^T$ is the Gram matrix. Thus the kernelized objective function is

$$J_k(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \ell((K\alpha)_i, y_i) + \lambda \alpha^T K \alpha.$$

You can also write out the objective function, in which case you can drop the $\log(y!)$ term, since it's independent of α.

(b) (1 point) Let w^* be a minimizer of $J(w)$, and let α^* be a minimizer of $J_k(\alpha)$. Give an expression relating w^* to α^*,

Solution: $w^* = \sum_{i=1}^n \alpha_i^* x_i = X^T \alpha^*$

(c) (1 point) Give a kernelized form of the prediction function $\hat{f}(x) = x^T w^*$.

Solution:

$$w^{*T} x = \left(\sum_{i=1}^{n} \alpha_i^* x_i\right) x = k_x^T \alpha^*$$

where $k_x^T = [\langle x_1, x\rangle, \cdots \langle x_n, x\rangle]$.

(d) (1 point) Which ONE of the following is the MOST accurate regarding kernel methods:

□ Good to use on very small data sets when unkernelized methods are overfitting.

- □ Good to use on very large data sets (e.g. millions or billions of data points) due to their efficiency.
- ■ **Good to use on medium-sized data sets when unkernelized methods are underfitting.**

(e) (1 point) __**F**__ **True or False**: Suppose $\varphi : \mathbf{R}^d \to \mathbf{R}^D$ is the feature map corresponding to our kernel function. Since $k(x, x') = \langle \varphi(x), \varphi(x') \rangle$ is an inner product in a D-dimensional space, kernel methods become infeasible when D is very large.