COMPSCI 671 Exam 2 (66 total pts.)

April 16th 2019

Name: NetId:

I will follow Duke academic standards _____ (signature required)

Thank you for following Duke academic standards!

(Turn over when exam starts)

1 Answer and Explain (21 points total, correct true or false label = 2 points, correct explanation = 1 point)

Answer the following questions, and write a short explanation for each statement.

1. Fill in the blank: The VC dimension of a _____ is the largest number of points that can it can shatter.

   Explanation (one sentence):

   Solution: Class of functions. VC dimension is a property of a class of functions, not a single function.

2. True or False: The VC dimension of the class of circles in $\mathbb{R}^2$, where a point is classified as a 1 if it is inside the circle and as a 0 if it is outside, is 4.

   Circle **True** or **False**

   Explanation:
   Solution: False. We can always shatter 3 points with a circle in 2d by arranging them as the vertices of an equilateral triangle and then drawing a circle around the positive examples, but we cannot find any such arrangement for 4 points.

3. True or False: Let $k(x, z)$ be a valid kernel mapping from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}$. If h is a polynomial function with positive coefficients, then $h(k(x, z))$ is also a valid kernel.

   Circle: **True** or **False**

   Explanation:

   Solution: True. We know this because, a) for positive α, β, $\alpha k(x, z) + \beta k(x, z)$ is also a valid kernel, and b) $\prod_{k=1}^{K} k(z, z)$ is also a valid kernel. Putting the two statements together, we can construct any polynomial function of k, which also has to be a valid kernel.

4. True or False: If the class of decision functions we are considering is finite, then we can construct a (non-vacuous) bound on the true risk for that class *without* using either VC dimension, Rademacher complexity, or covering numbers.

   Circle: **True** or **False**

   Explanation:
   Solution: True. We can use the Occham's Razor bound, which uses Hoeffding's inequality and the union bound.

5. True or False: Let $k(x, z)$ be a valid kernel, and let H_k be its corresponding Reproducing Kernel Hilbert Space. The representer theorem states that, for any function $\ell : \mathbb{R}^2 \mapsto \mathbb{R}$, if

$$f^* \in \operatorname*{argmin}_f \sum_{i=1}^{n} \ell(f(x_i), y_i) + ||f||^2_{H_k},$$

   then $f^* = \prod_{i=1}^{n} \alpha_i k(x_i, \cdot)$.

   Circle: **True** or **False**

   Explanation:
   Solution: False. The representer theorem states that, if $f^* \in \operatorname{argmin} \sum_{i=1}^{n} \ell(f(x_i), y_i) + ||f||^2_{H_k}$, then $f^* = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$.

6. Name one algorithm that has both a natural frequentist interpretation and a Bayesian interpretation:

   As an explanation, state what kind of Bayesian prior is used for this algorithm:

   Solution: Ridge regression, least squares, and logistic regression are all solutions to this problem. Ridge regression has a normal prior, whereas the others have uniform priors.

7. Consider the Multi-Armed-Bandit problem with m arms and let $\hat{X}_j(t)$ be the estimated reward for playing arm j at iteration t. At iteration t, the ϵ-greedy algorithm plays one arm uniformly at random with probability ϵ_t, and with probability $1 - \epsilon_t$ the arm j such that $\hat{X}_j(t-1) \leq \hat{X}_i(t-1)$, for all arms $i = 1 \ldots, m$.

   Circle: **True** or **False**

   Explanation:

   Solution: False. The ϵ-greedy algorithm plays the arm with the greatest estimated reward with probability $1-\epsilon_t$, i.e., arm j such that: $\hat{X}_j(t-1) \geq \hat{X}_i(t-1)$, for all arms $i = 1 \ldots, m$.

2 Optimality (11 points total)

a) **(7 points)** For each of the following algorithms, state whether it is true or false that they are guaranteed to produce a globally optimal solution to an optimization problem. You do not need to provide an explanation, just circle the answer.

- Support Vector Classification with Gaussian Kernels (Circle one: True / False)
- Expectation-Maximization (Circle one: True / False)
- K-Means (Circle one: True / False)
- Lasso (Circle one: True / False)
- Logistic Regression (Circle one: True / False)
- Boosted Stumps (Circle one: True / False)
- CART (Circle one: True / False)

Solution: T,F,F,T,T,T,F

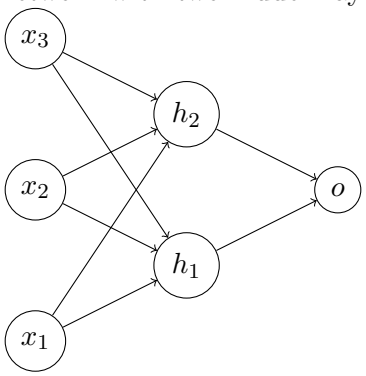
b) **(4 points)** Write down the optimization problem one would solve when performing logistic regression for binary classification.

Solution: $\min_\lambda \sum_i \log(1 + \exp(-y_i f_\lambda(x_i))$ where $f_\lambda(x_i) = \sum_j \lambda_j x_{ij}$.

3 Neural Network (12 points)

Consider the neural network in Figure 1: This network has three inputs: $\mathbf{x} = [x_1, x_2, x_3]^T$ and two

Figure 1: A Neural Network with two hidden layers and three inputs.



hidden layers: $\mathbf{h} = [h_1(\mathbf{x}), h_2(\mathbf{x})]^T$. The hidden layers, as well as the output layer have a sigmoid activation:

$$h_k(\mathbf{x}) = \sigma(\mathbf{w}_k^T \mathbf{x}), \text{ and } o = \sigma(\mathbf{v}^T \mathbf{h}),$$

where $k = 1, 2$ and,

$$\sigma(a) = \frac{1}{1 + e^{-a}},$$

$\mathbf{w}_k \in \mathbb{R}^3$ is a vector of three weights, one for each of the three inputs, and $\mathbf{v} \in \mathbb{R}^2$ is a vector of two weights for the hidden layers.

a) (5 points) Write the prediction function for the network $\hat{y} = f(\mathbf{x})$ in terms of $\mathbf{w}_1, \mathbf{w}_2, \mathbf{v}$, and σ.

solution: $f(\mathbf{x}) = \sigma(\mathbf{v}^T[\sigma(\mathbf{w}_1\mathbf{x}), \sigma(\mathbf{w}_2\mathbf{x})]^T)$

b) (5 points) The error term for the network on one observation, y, is $\epsilon = \frac{1}{2}(y - f(\mathbf{x}))^2 = \frac{1}{2}(y - o)^2$. Recall that $\frac{d\sigma(z)}{dz} = \sigma(z)(1 - \sigma(z))$, and use this to write the gradients of the error with respect to $\mathbf{v}$ and $\mathbf{w}_1, \mathbf{w}_2$ in terms of $\sigma, \mathbf{v}, \mathbf{w}$, and $\mathbf{x}$. It will possibly be easier if you use dot product notation rather than summation notation.

solution:

$$\frac{\partial \epsilon}{\partial \mathbf{v}} = \frac{\partial \epsilon}{\partial o}\frac{\partial o}{\partial \mathbf{v}} = -\sigma(\mathbf{v}^T\mathbf{h})(1 - \sigma(\mathbf{v}^T\mathbf{h}))\mathbf{h}.$$

$$\frac{\partial \epsilon}{\partial \mathbf{w}_1} = \frac{\partial \epsilon}{\partial o}\frac{\partial o}{\partial \mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{w}_1} = -\sigma(\mathbf{v}^T\mathbf{h})(1 - \sigma(\mathbf{v}^T\mathbf{h}))\mathbf{v}^T[\sigma(\mathbf{w}_1^T\mathbf{x})(1 - \sigma(\mathbf{w}_1^T\mathbf{x}))\mathbf{x}^T, 0]$$

$$\frac{\partial \epsilon}{\partial \mathbf{w}_1} = \frac{\partial \epsilon}{\partial o}\frac{\partial o}{\partial \mathbf{h}}\frac{\partial \mathbf{h}}{\partial \mathbf{w}_2} = -\sigma(\mathbf{v}^T\mathbf{h})(1 - \sigma(\mathbf{v}^T\mathbf{h}))\mathbf{v}^T[0, \sigma(\mathbf{w}_2^T\mathbf{x})(1 - \sigma(\mathbf{w}_2^T\mathbf{x}))\mathbf{x}^T].$$

c) (2 points) What are the benefits and drawbacks of adding more hidden layers to a network? Provide a list of at least 1 possible benefit and 2 possible drawbacks.

solution: Benefits: more parameters could allow better fitting. Drawbacks: A network with too many layers might overfit the training data. It also might be harder to train.

4 Support Vector Machines (15 Points)

Let $\{(x_i, y_i)\}_{i=1}^n$ be a set of n training pairs of feature vectors and labels. Let the labels $y_i \in \mathbb{R}$ be real numbers. We would like to use SVM to solve a **regression** problem in this case. To do so we adopt the following formulation of the SVM optimization problem:

$$\min_{\mathbf{w},\xi,\xi^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

subject to:

$$y_i - \mathbf{w}^T\mathbf{x}_i - b \leq \xi_i \qquad \forall i$$

$$\mathbf{w}^T\mathbf{x}_i + b - y_i \leq \xi_i^* \qquad \forall i.$$

This is a simplified version of the SVM problem for regression.

a) (5 Points) Give the Lagrangian of the problem.

Solution:

$$\mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*) = \min_{\mathbf{w},\xi,\xi^*} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) + \sum_{i=1}^{n}\alpha_i(y_i - \mathbf{w}^T\mathbf{x}_i - b - \xi_i) + \sum_{i=1}^{n}\alpha_i^*(\mathbf{w}^T\mathbf{x}_i + b - y_i - \xi_i^*)$$

b) (10 Points) Write down the KKT conditions (lagrangian stationarity, primal feasibility, dual feasibility, complementary slackness) of the problem. You do not try to solve the dual by combining equations, as we will only grade the KKT conditions themselves.

Solution: Lagrangian stationarity:

$$\frac{\delta \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{\delta \mathbf{w}} = \mathbf{w} + \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)\mathbf{x}_i = 0$$
$$\frac{\delta \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{\delta b} = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) = 0$$
$$\frac{\delta \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{\delta \xi_i} = C - \alpha_i = 0$$
$$\frac{\delta \mathcal{L}(\mathbf{w}, b, \xi_i, \xi_i^*)}{\delta \xi_i^*} = C - \alpha_i^* = 0.$$

Primal feasibility:

$$y_i - \mathbf{w}^T \mathbf{x}_i - b - \xi_i \leq 0 \qquad \forall i$$
$$\mathbf{w}^T \mathbf{x}_i + b - y_i - \xi_i^* \leq 0 \qquad \forall i$$

Dual feasibility:

$$\alpha_i \geq 0 \qquad \forall i$$
$$\alpha_i^* \geq 0 \qquad \forall i$$

Complementary slackness:

$$\alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i - b - \xi_i) = 0 \qquad \forall i$$
$$\alpha_i^* (\mathbf{w}^T \mathbf{x}_i + b - y_i - \xi_i^*) = 0 \qquad \forall i$$

5 Mixture Models (7 points)

We have data on the failure time of a set of K machines, $\{y_1, \ldots, y_n\}$, where $y_i \in \mathbb{R}^+$. The observed quantity y_i denotes how long after startup a failure was recorded. Unfortunately, we do not know which of the observed failure times refer to which ones of the K machines. To deal with this problem, we model the data as a mixture of K exponential distributions:

$$\Pr(\{Y_i = y_i\}_i | \boldsymbol{\lambda}, \boldsymbol{\pi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_k \lambda_k e^{-\lambda_k y_i}, \quad 0 \leq \pi_k \leq 1, \quad \sum_{k=1}^{K} \pi_k = 1,$$

where $\lambda e^{-\lambda y}$ is the pdf of the exponential distribution, and π_k is the probability of the observation coming from machine k, that is:

$$\Pr(Z_i = k) = \pi_k,$$

and where we had introduced latent variables $\{Z_1, \ldots, Z_n\}$, $Z_i \in \{1, \ldots, k\}$ representing whether observation i belongs to machine k.

a) (3 points) Write down the probability of observing y_i, given that observation i belongs to machine k, that is $\Pr(Y_i = y_i | Z_i = k, \boldsymbol{\lambda}, \boldsymbol{\pi})$.

solution:

$$\Pr(Y_i = y_i | Z_i = k, \lambda_k) = \lambda_k e^{-\lambda_k y_i}.$$

b) (3 points) Write down the posterior probability of observation i coming from machine k, that is, $\Pr(Z_i = k | y_i, \boldsymbol{\lambda}, \boldsymbol{\pi})$.

solution:

$$\Pr(Z_i = k | y_i, \boldsymbol{\lambda}, \boldsymbol{\pi}) = \frac{\pi_k \lambda_k e^{-\lambda_k y_i}}{\sum_{k=1}^{K} \pi_k \lambda_k e^{-\lambda_k y_i}}.$$