Machine Learning 2019-20
Final Exam

18 December 2019

Name: .....

Question 1: 1.5 points Consider the Perceptron learning algorithm that sequentially updates a parameter vector using the rule $\mathbf{w}_{t+1} = \mathbf{w}_t + y_{n(t)}\mathbf{x}_{n(t)}$, where $(\mathbf{x}_{n(t)}, y_{n(t)})$ is a data point that is misclassified by the linear classifier specified by $\mathbf{w}_t$. Show the following two properties of the update:

- $y_{n(t)}\mathbf{w}_t^\top \mathbf{x}_{n(t)} < 0$.
- $y_{n(t)}\mathbf{w}_{t+1}^\top \mathbf{x}_{n(t)} > y_{n(t)}\mathbf{w}_t^\top \mathbf{x}_{n(t)}$.

Question 2: 1.5 points Define the following components of a general supervised learning system: (i) input and output spaces, (ii) hypothesis set, (iii) loss function (or error measure). How are these components chosen for linear regression and logistic regression?

Question 3: [1 point] Consider a regression problem with the target function $f(x) = 3+x-6x^3+2x^6$, and N data points of the form $y_t = f(x_t) + \varepsilon_t$, where ε_t is Gaussian noise with zero mean and variance σ^2, and x_t is distributed uniformly on $[0, 1]$. We are given two hypothesis sets: the set of second-order polynomials $\mathcal{H}_2$ and the set of 10th order polynomials $\mathcal{H}_{10}$. Which hypothesis set would you use when (i) $\sigma^2 = 0$ and $N = 100$, (ii) $\sigma^2 = 1$ and $N = 20$. Explain why!

Question 4: [1 point] What are the relative advantages and disadvantages of using the Perceptron learning algorithm and soft-margin SVMs for linear classification?

Question 5: [1 point] Consider a data set with 2-dimensional inputs $\mathbf{x}$ distributed uniformly over the unit cube $[0, 1]^2$ and with labels y generated as follows: $y = +1$ if $x_1 > \frac{1}{2}$ and $x_2 > \frac{1}{2}$, and $y = -1$ otherwise. Which of these techniques do you think is least suitable for learning this target function and why: (i) linear SVM (ii) SVM with Gaussian kernel (iii) decision trees, or (iv) neural networks?

Name: .....

Question 6: 1.5 points Describe the procedure of V-fold cross-validation and briefly explain its advantage over choosing the hyperparameters on the training set.

Question 7: 1.5 points

A data point $\mathbf{x}$ can be modeled using a probabilistic model $p(\mathbf{x}|\boldsymbol{\theta})$ that depends on some parameters $\boldsymbol{\theta}$. In light of some dataset $\mathcal{D}$ that comprises many data points,

1. Write down the formula describing how a Bayesian would learn the model parameters $\boldsymbol{\theta}$.
2. Write down the differences between the previous formula and what a maximum likelihood estimator would do.
3. Write down the formula used by a Bayesian to model predictions of a new data point $\mathbf{y}$.
4. How does the previous predictions differ from the prediction based on the maximum likelihood estimator?

Question 8: [1 point] A standard XOR gate is given by the following table

A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

1. If we observe that the output of the XOR gate is 0, what can we say about A and B?

Consider a 'soft' version of the probabilistic XOR gate given by the following table

A	B	P(C=1|A,B)
0	0	0.1
0	1	0.99
1	0	0.8
1	1	0.25

2. Assuming additionally that $A \perp\!\!\!\perp B$, and $p(A = 1) = 0.65, p(B = 1) = 0.77$. What is $p(A = 1|C = 0)$?
3. Write down the corresponding graphical model