Machine Learning 2018-19
Final Exam

12 December 2018

Name: .....

Question 1: 2 points In linear regression, the *hat matrix* $\mathrm{H} = \mathrm{X}(\mathrm{X}^\top\mathrm{X})^{-1}\mathrm{X}^\top$ puts the "hat" on the true output $\mathbf{y}$, i.e. the estimated output equals $\hat{\mathbf{y}} = \mathrm{H}\mathbf{y}$. Show that a) $\mathrm{H}^\top = \mathrm{H}$; b) $\mathrm{H}^2 = \mathrm{H}$; c) $(\mathrm{I} - \mathrm{H})^2 = (\mathrm{I} - \mathrm{H})$. Hints: For any multiplicable matrices A and B, $(\mathrm{AB})^\top = \mathrm{B}^\top\mathrm{A}^\top$, and for any square matrix C, $(\mathrm{C}^{-1})^\top = (\mathrm{C}^\top)^{-1}$.

Question 2: 2 point Which are the components of supervised learning? Which of these components can be controlled by a machine learning practitioner?

Question 3: 1 point What is overfitting? Name three factors that increase the amount of overfitting.

Question 4: 1 point What is regularization? Explain how regularization is defined theoretically, and how regularization is usually implemented in practice. Why is there a difference?

Question 5: 1 point In decision tree learning, how is information gain defined? Describe the standard algorithm for learning decision trees.

Question 6: 1 point Explain how recurrent neural networks differ from feedforward neural networks. Describe one advantage and one disadvantage of recurrent neural networks, compared to feedforward neural networks.

Name: .....

Question 7: 1 point Write down the Bayes Theorem. Describe two possible uses of this theorem in Machine Learning. For each use, describe precisely each term in the formula.

Question 8: 2 point

A belief network models the relation between the variables oil (oil), inflation (inf), economy health (eh), British petroleum stock price (bp), and retailer stock price (rt). Each variable takes the states $\{low, high\}$ except for bp which has states $\{low, high, normal\}$. The Bayesian network model for these variables has tables

$$p(eh = low) = 0.2$$

$$p(bp = low|oil = low) = 0.9 \qquad p(bp = normal|oil = low) = 0.1$$

$$p(bp = low|oil = high) = 0.1 \qquad p(bp = normal|oil = high) = 0.4$$

$$p(oil = low|eh = low) = 0.9 \qquad p(oil = low|eh = high) = 0.05$$

$$p(rt = low|inf = low, eh = low) = 0.9 \qquad p(rt = low|inf = low, eh = high) = 0.1$$

$$p(rt = low|inf = high, eh = low) = 0.1 \qquad p(rt = low|inf = high, eh = high) = 0.01$$

$$p(inf = low|oil = low, eh = low) = 0.9 \qquad p(inf = low|oil = low, eh = high) = 0.1$$

$$p(inf = low|oil = high, eh = low) = 0.1 \qquad p(inf = low|oil = high, eh = high) = 0.01$$

1. **(1pt)** Draw the corresponding Bayesian network for this distribution.
2. **(0.5pt)** Use the D-Separation algorithm to argue whether the following conditional independences are satisfied or not.
   (a) Is eh independent of bp if no evidence is provided? Why?
   (b) Is eh independent of bp if we observe that the oil is $high$? Why?
   (c) Is rt independent of bp if we observe that eh is low? Why?
3. **(0.5pt)** Given that the British stock price is normal ($bp = normal$) and the retailer stock price is high ($rt = high$), what is the probability that inflation is high ($inf = high$)?

(answers on next page)

Name: .....

Question 9: [1 point] In the context of discounted Markov decision processes, define the value function V^π associated with a deterministic policy $\pi : \mathcal{X} \to \mathcal{A}$.
Hint: give either the basic definition involving an infinite sum, or the Bellman equations.

Question 10: [1 point] The Policy Iteration algorithm for discounted Markov decision processes iteratively computes a sequence of value functions and policies $V_1, \pi_1, V_2, \pi_2 \ldots, V_k, \pi_k$. In a given iteration k of the algorithm, what is the relation between V_k and π_k (i.e., the policy computed right after computing V_k)?

Question 11: [2 point] Describe the SARSA and the Q-learning algorithms for discounted Markov decision processes. What are the respective behavior and target policies of the two algorithms?