Qualification Exam Question

1 Statistical Models and Methods

1.1 Core

1. **Cross-validation** We would like to perform k-fold cross-validation. What should k be? Discuss the pros and cons of large or small values of k.

2. **Bayes classifier**

   (a) Write down the Bayes classifier $f : X \to Y$ (the classifier that minimizes the expected loss $E(L(Y, f(X))))$ for binary classification $Y \in \{-1, +1\}$ with non 0-1 loss (a is the loss for falsely predicting negative and b is the loss for falsely predicting positive). Simplify the classification rule as much as you can.

   (b) If $P(X|Y = y)$ is a multivariate Gaussian and assuming the 0/1 loss, write the Bayes classifier as $f(X) = \text{sign}(h(X))$ and simplify h as much as possible. What is the geometric shape of the decision boundary?

   (c) Repeat (b) when the two Gaussians have identical covariance matrices. What is the geometric shape of the decision boundary?

   (d) Repeat (b) when the two Gaussians have covariance matrix that equals the identity matrix. Describe the geometric shape of the decision boundary as much as possible.

3. **Multiclass classification**

   Multiclass classification tries to assign one of several class labels (rather than binary labels) to an object. Can you give two ways which use binary classifiers to solve multiclass classification problem? What are the pros and cons of these different methods (eg. in terms of computional complexity or the applicability of the method)? Besides using binary classifiers, do you have any other idea on how to build a multiclass classifier?

1.2 Methods and Models

1. **Kernel methods**

   Consider two machine learning models for 2-class classification. The first is a support vector machine with Gaussian kernel. The second is kernel discriminant analysis (a Bayes classifier with a kernel density estimator for each class), where the bandwidth may vary for each dimension, and possibly also for each data point. Which is the more expressive, or powerful model? Compare and discuss the pros and cons of each.

2. **Bayes Rule**

   Let $\phi(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$ denote the density of a random variable y with a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Suppose that we have three related random variables, X, Y and Z,

- Random variable X has a Gaussian distribution $\mathcal{N}(0, \sigma^2)$;
- Given random variable $X = x$, random variable Y has a Gaussian distribution $\mathcal{N}(x, \sigma^2)$;
- Given random variable $Y = y$, random variable Z is a mixture of two Gaussians with density

$$p(z|Y = y) = (1 - \alpha)\phi(z; 0, \sigma^2) + \alpha\phi(z; y, \sigma^2). \tag{1}$$

- Conditioned on random variable Y, random variables X and Z are independent

Given n i.i.d. sample $z^1, \ldots, z^n$ from the mixture density (1), answer the following questions

(a) If $n = 2$, derive the posterior distribution of X conditioned on $(z^1, \ldots, z^n)$ exact upto a scalar difference.

(b) If $n = 10$ (or in general when n is large), what is the computational problem associated with computing the poterior distribution of X?

(c) Propose approximation algorithms to deal with the computational problem when n is large.

3. **Dependent noise model**

Let $X_1, \ldots, X_n$ be n determinations of a physical constant θ. Consider the model,

$$X_i = \theta + e_i, \quad i = 1, \ldots, n$$

and assume

$$e_i = \alpha e_{i-1} + \beta e_{i-2} + \epsilon_i, \quad i = 1, \ldots, n, \quad e_0 = 0, e_{-1} = 0$$

with ϵ_i's iid standard normal, and α and β are known constant. What is the maximum likelihood estimate of θ? Carefully justify each step of your derivation/calculation.

2 Learning Theory

1. **VCdimension**

(a) What is the VC-dimension of axis-parallel rectangles in R^3? Specifically, a legal target function is specified by three intervals $[x_{min}, x_{max}]$, $[y_{min}, y_{max}]$, and $[z_{min}, z_{max}]$, and classifies an example (x, y, z) as positive iff $x \in [x_{min}, x_{max}]$, $y \in [y_{min}, y_{max}]$, and $z \in [z_{min}, z_{max}]$. (b) Describe the importance of VC-dimension for Machine Learning.

2. **Mistake-bound model.**

(a) k-CNF is the class of Conjunctive Normal Form formulas in which each clause has size at most k. E.g., $x_4 \wedge (x_1 \vee x_2) \wedge (x_2 \vee \bar{x}_3 \vee x_5)$ is a 3-CNF. Give an algorithm to learn 5-CNF formulas over n boolean features in the mistake-bound model. Your algorithm should run in polynomial-time per example (so the "halving algorithm" is not allowed). How many mistakes does it make at most? (b) What is the relationship between the mistake bound model and the PAC learning model?

3. **Consistency Problem for 2-term DNF formulas** (a) Prove that the consistency problem for 2-term DNF formulas is NP-hard. (b) Is the class of 2-term DNF formulas PAC-learnable? Explain why or why not.

3 Decision Processes

The theme is scalability, and you aren't getting out of it.

1. **Scaling up reinforcement learning**

   Machine learning algorithms have traditionally had difficulty scaling to large problems. In classification and traditional supervised learning this problem arises with data that exist in very high dimensional spaces or when there are many data points for computing, for example, estimates of conditional densities. In reinforcement learning this is also the case, arising when, for example, there are many, many states or when actions are at a very low level of abstraction.

   - Typical approaches to addressing such problems in RL include function approximation and problem decomposition. Compare and contrast these two approaches. What problems of scale do these approaches address? What are their strengths and weaknesses? Are they orthogonal approaches? Can they work well together?
   - What are the differences between hierarchical and modular reinforcement learning? Explain both the theoretical and practical limits of these approaches.

2. **Learning from demonstrations**

   Machine learning algorithms have traditionally had difficulty scaling to large problems. In classification and traditional supervised learning this problem arises with data that exist in very high dimensional spaces or when there are many data points for computing, for example, estimates of conditional densities. In reinforcement learning this is also the case, arising when, for example, there are many, many states or when actions are at a very low level of abstraction.

   Imagine that we want to leverage domain knowledge from humans in order attack this problem of scalability. One mechanism we might use is Learning from Demonstration where humans demonstrate correct behavior; however, complex tasks can require more examples of complete behavior than is practical to obtain. Given that you will only be able to extract so much time from your human teachers, what are at least two ways you might still take advantage of their ability to give demonstrations, even for complex tasks? For each proposed method, describe strengths and possible pitfalls.

3. **Learning with Options**

   Machine learning algorithms have traditionally had difficulty scaling to large problems. In classification and traditional supervised learning this problem arises with data that exist in very high dimensional spaces or when there are many data points for computing, for example, estimates of conditional densities. In reinforcement learning this is also the case, arising when, for example, there are many, many states or when actions are at a very low level of abstraction.

   One mechanism for addressing such concerns in RL is to use so-called options. Options are a mechanism for incorporating temporally-extended actions into the RL framework.

   - Formally define an option.
   - What are the advantages and limits of options? Be specific.
   - Describe at least two ways one might automate the process of creating options.
   - Although options are defined in a very specific way, would you argue that different options might serve different purposes? If so, do these different kinds of options have

identifiably different properties? If you believe that different options do not serve different purposes, argue for that position as well.