Take-Home Mid-Term Examination. Turn in scanned/images of examples. You can perform additional work and attach. This exam has 14 questions, for a total of 105 points.

1. Given a sample of data $\{x_i\}_{i=1}^n$ where each data point is a vector $x_1, \ldots, x_n \in \mathbb{R}^d$, let $J(p)$ denote a cost function in terms of the dimensionality of a optimally trained linear auto-encoder,

$$J^*(p) = \min_{\mathbf{U} \in \mathbb{R}^{p \times d}, \mathbf{V} \in \mathbb{R}^{(p+1) \times d}} J(\mathbf{U}, \mathbf{V}) \tag{1}$$

$$J(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{n} \| [(\mathbf{U}\mathbf{x}_i)^\top, 1]\mathbf{V} - \mathbf{x}_i^\top \|_2^2, \tag{2}$$

where $\mathrm{Encode}(\mathbf{x}) = \mathbf{U}\mathbf{x} = \mathbf{z} \in \mathbb{R}^p, \mathrm{Decode}(\mathbf{z}) = \mathbf{V}^\top [\mathbf{z}^\top, 1]^\top = \hat{\mathbf{x}} \in \mathbb{R}^d$.

(a) (5 points) Should $J^*(p)$ be used directly to choose the dimensionality? (Consider that you want to apply the auto-encoder to new data points to denoise them.) If not, how would you propose to select the best dimensionality p?

(b) (5 points) For what values of p should $J^*(p) = 0$?

(c) (5 points) For $p = 0$, what vector of values should the only column of the optimal solution $\mathbf{V}^*$ have?

(d) (5 points) If $p = 1$ what should the first column vector of $\mathbf{V}^*$ be?

2. (10 points) Decision trees often use an entropy-like quantity to choose the feature and threshold (for continue valued features) at each internal node. If you used Shannon's entropy, describe this selection process in terms of information theoretic quantities.

3. (10 points) Describe two distinct approaches for scaling linear models in the case that the number of features is much larger than the number of training instances. (Hint: one is based on the form of the optimization problem, and the other is based on randomization).

4. (5 points) Describe how a kernel regression with a radial basis function kernel has a linear parameterization and provide a non-linear prediction.

5. (5 points) Describe at two advantages of kernel SVM over a kNN classifier, even if both rely on the Euclidean distance (input to the Gaussian kernel in the former case).

6. (5 points) Describe an advantage and disadvantage of kernel SVM compared to a linear SVM for classification. You may want to discuss examples where each would fail.

7. (5 points) What is the goal of random Fourier feature embedding?

8. (5 points) Describe how k-means can be used in the context of the Nyström method.

9. (5 points) What are at least two approaches (and corresponding assumptions) in a multi-task framework that can be applied to regularize a set of linear-models?

10. (5 points) Name a non-parametric alternative for an unpaired t-test.

11. (10 points) If you were testing for independence between two discrete-valued random variables, describe how you would generate a surrogate distribution for a test-statistic under the null hypothesis.

12. (5 points) If two univariate, continuous-valued random variables have the same mean and variance can we say that the distributions are equal? If not provide a counter example.

13. (5 points) If two univariate, continuous-valued random variables have the same cumulative distribution function, can we say that the distributions are equal? If not provide a counter example.

14. (10 points) What are two approaches for approximating the Wasserstein distance (optimal transport) for large samples?