Final Exam Review

10-701: Introduction to Machine Learning

4/26/2024

1 Unsupervised Learning

1.1 Clustering

1. For each of the **True or False** questions below, select the correct answer and briefly justify your selection in 1-2 concise sentences.

   (a) For a fixed dataset and k, the k-means algorithm will always produce the same result if the initial centers are the same.

   ○ True

   ○ False

   (b) The k-means algorithm will always converge to the globally optimal solution.

   ○ True

   ○ False

(c) In the k-means algorithm, the objective function's value can increase, decrease or stay the same after each iteration.

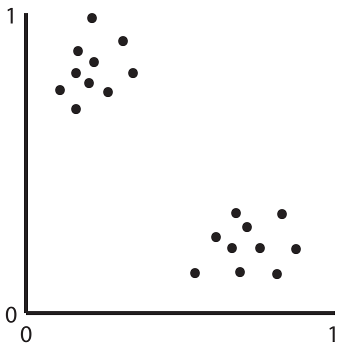
○ True

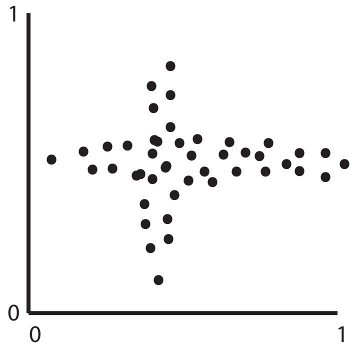
○ False

2. **Short answer:** In the setting of expectation-maximization for Gaussian mixture models, define the marginal log-likelihood (either in words or mathematically) and descrbie why we optimize the marginal log-likelihood instead of the complete log-likelihood of the dataset.

1.2 Dimensionality Reduction

1. **Drawing:** Consider the following two datasets. On the figures below, draw arrows from the mean of the data to denote the direction and relative magnitudes of the principal components.





2. Given a dataset $\mathcal{D}$ consisting of N data points and D features, suppose you use PCA to project the dataset down to $d < D$ dimensions: let E be the squared reconstruction error of this projection.

   (a) **Select one:** Now suppose that you add an extra data point to $\mathcal{D}$ so that $\mathcal{D}'$ consists of $N+1$ data points and D features. You once again use PCA to project $\mathcal{D}'$ to $d < D$ dimensions: let E' be the squared reconstruction error of this new projection. How do E and E' relate to one another?

   ○ $E < E'$

   ○ $E \leq E'$

   ○ $E = E'$

   ○ $E \geq E'$

   ○ $E > E'$

   (b) **Select one:** Now suppose that you use PCA to project the original dataset $\mathcal{D}$ down to $(d+1) < D$ dimensions instead of d dimensions: let E' be the squared reconstruction error of this new projection. How do E and E' relate to one another?

   ○ $E < E'$

   ○ $E \leq E'$

   ○ $E = E'$

   ○ $E \geq E'$

   ○ $E > E'$

3. **Select all that apply:** Recall from lecture that autoencoders are trained by minimizing the reconstruction error between the inputs $\mathbf{x}$ and the corresponding outputs $\mathbf{x}'$. Given a dataset, $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$, which of the following are rational alternative objective functions for training autoencoders with backpropagation?

   □ $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}^{(i)'}\|_2^2$

   □ $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}^{(i)'} - \mathbf{x}^{(i)}\|_1$

   □ $\underset{i}{max}\|\mathbf{x}^{(i)'} - \mathbf{x}^{(i)}\|_2^2$

   □ $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}^{(i)'} - \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}^{(i)'}\|_2^2$

   □ None of the above.

2 Reinforcement Learning

1. Formulate the task of deciding how long to study for each of our two 10-701 exams this semester as a single reinforcement learning problem. Specifically, briefly describe in words the following components:

   (a) **Short Answer:** The state space $\mathcal{S}$.

   (b) **Short Answer:** The action space $\mathcal{A}$; make sure to specify what actions can be taken in any given state.

   (c) **Short Answer:** The reward function or distribution.

   (d) **Short Answer:** The transition function or distribution.

2. **Select one:** Which algorithm is *most* appropriate for solving the problem you defined above? Briefly justify your answer in 2-3 concise sentences.

   ○ Value iteration

   ○ Policy iteration

   ○ Q-learning

   ○ Deep Q-learning

3. **Math:** In class we taught $\pi(s)$ as the action taken in state s under deterministic policy π. Now we consider the case of a stochastic policy π, such that $\pi(a \mid s)$ is the probability of taking action a in state s under stochastic policy π. If the current state is s_t, write the expectation of r_{t+1} in terms of $\pi(a \mid s)$ and $p(s', r \mid s, a)$, the probability of transitioning to state s' with reward r, from state s and action a.

4. **Select all that apply:** Which of the following are *necessary* conditions for Q-learning to converge to the optimal Q values?

   □ For every state, every valid action is taken at least once.

   □ The discount factor is *strictly* between 0 and 1.

   □ All rewards and all initial Q values are finite.

   □ The learning rate is constant.

   □ None of the above.

3 Pretraining

1. **Short answer:** In 2-3 concise sentences, describe the relationship between unsupervised pretraining and the autoencoder architecture.

2. **Select all that apply:** Which of the following are issues or shortcomings with policy gradient methods such as the likelihood ratio method presented in lecture?

   □ The learned policies must be deterministic.

   □ The number of sampled trajectories needed to estimate the objective function may be very large.

   □ Policy gradient methods are incompatible with contiuous state and action spaces.

   □ Trajectories are sampled from the current policy, which can be suboptimal or unsafe.

   □ None of the above.

3. **Select one:** For a fixed large language model and natural language task, how do few-, one- and zero-shot accuracy tend to relate to one another and how is this affected by model size?

   ○ Few-shot accuracy, one-shot accuracy and zero-shot accuracy are roughly equivalent and stay roughly constant across all model sizes.

   ○ Few-shot accuracy, one-shot accuracy and zero-shot accuracy are roughly equivalent and tend to increase with increasing model size.

   ○ Few-shot accuracy is greater than one-shot, which is greater than zero-shot and the differences are roughly constant across different sized models.

   ○ Few-shot accuracy is greater than one-shot, which is greater than zero-shot and the differences tend to increase with increasing model size.

4 Algorithmic Bias

1. **Numerical answer:** Suppose you have binary classification dataset where 20% of the data points have label +1 and the remaining 80% have label −1. What are the precision, recall, and F1-score of a classifier that always predicts +1?

2. **Select all that apply:** Suppose you have a classifier h that is always 100% accurate at some binary classification task. Furthermore, suppose that there is some protected attribute, A, and the percentage of positive labels is *constant* across different values of A. In this setting, which of the following definitions of algorithmic fairness is satisfied by h?

   □ Independence

   □ Separation

   □ Sufficiency

   □ None of the above.

3. **Select all that apply:** Suppose you have a classifier h that always predicts a random label for some binary classification task. Furthermore, suppose that there is some protected attribute, A, and the percentage of positive labels varies across different values of A. In this setting, which of the following definitions of algorithmic fairness is satisfied by h?

   □ Independence

   □ Separation

   □ Sufficiency

   □ None of the above.

5 Learning Theory

1. **Fill in the Blanks:** Complete the following sentence by circling one option in each square:

   In order to prove that the VC-dimension of a hypothesis set $\mathcal{H}$ is D, you must show that $\mathcal{H}$ can / cannot shatter any set / some set / multiple sets of D data points and can / cannot shatter any set / some set / multiple sets of $D+1$ data points.

2. **Math:** For an arbitrary finite hypothesis set $\mathcal{H}$, provide an upper bound on the VC-dimension of $\mathcal{H}$ in terms of $|\mathcal{H}|$.

3. Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{+1, -1\}$. Let

$$H = \{h_{a_1,a_2} \mid a_1, a_2 \in \mathbb{R}^2, h_{a_1,a_2}(x_1, x_2) = +1 \text{ iff } x_1 \leq a_1 \text{ and } x_2 \leq a_2\},$$

   be the hypothesis class corresponding to positive "quarter planes" in $\mathbb{R}^2$

   (a) **Math:** Show that the VC dimension of H is at least 2.

(b) **Math:** Show that the VC dimension of H is at most 2.

(c) **Math:** Verify the Sauer-Shelah lemma for H by showing that the number of possible distinct labellings produced by H on m points is $O(m^2)$.

6 Ensemble Methods

6.1 Bagging

1. **Short Answers:** Random forests reduce the variance of single decision trees by introducing randomness at different stages of the algorithm: what are the places where we introduce this randomness? For each assertion below, indicate whether it is true or false; if you select false, justify your answer in 1-2 concise sentences.

   (a) **Bootstrap Aggregation:** We choose N random examples without replacement from the dataset every time we train a decision tree in the forest. Doing so ensures that every tree looks at a different set of examples and thus, the forest as a whole will not overfit to the dataset.

   (b) **Bagging:** Take N random examples with replacement from the dataset every time we train a decision tree in the forest. Doing so and then combining the hypothesis of all trees (by taking majority) reduces variance while still holding the trends and statistical properties of the original dataset.

   (c) **Feature Split Randomization:** Every tree starts with a random subset of features and uses ID3 to split them. This ensures that all trees are not dependent on the same set of features and the forest is robust

   (d) **Hypothesis Combination/Aggregation:** We take the majority vote from a random subset of the decision trees at the end. This means that not all decision trees contribute to the final prediction, so the aggregated model is resilient to some trees having high variance.

2. **Select one:** In an effort to reduce the training error of each individual tree in a random forest, you decide to not perform split-feature randomization and allow every split to consider all of the features. How would you expect this to impact the generalization error of the random forest and why?

   ○ The generalization error will decrease as each individual decision tree will have lower training error.

   ○ The generalization error will decrease as each individual decision tree will make the same splits.

   ○ The generalization error will increase as the depth of each individual decision tree will tend to increase.

   ○ The generalization error will increase as the individual decision trees will become more correlated.

6.2 Boosting

1. You are developing a new boosting algorithm based off of AdaBoost and decide to use the following update rule for the weights:

$$\omega_t^{(i)} = \frac{\omega_{t-1}^{(i)} \alpha_t^{-\frac{1}{2} y^{(i)} h_t(\mathbf{x}^{(i)})}}{Z_t}$$

   (a) **Math:** Derive an expression for the normalization constant, Z_t, as a function of α_t and ϵ_t, the weighted training error.

(b) **Math:** Using your answer to part (a), compute the value of α_t that minimizes the normalization constant, Z_t. Express your answer as a function of ϵ_t.

7 SVMs

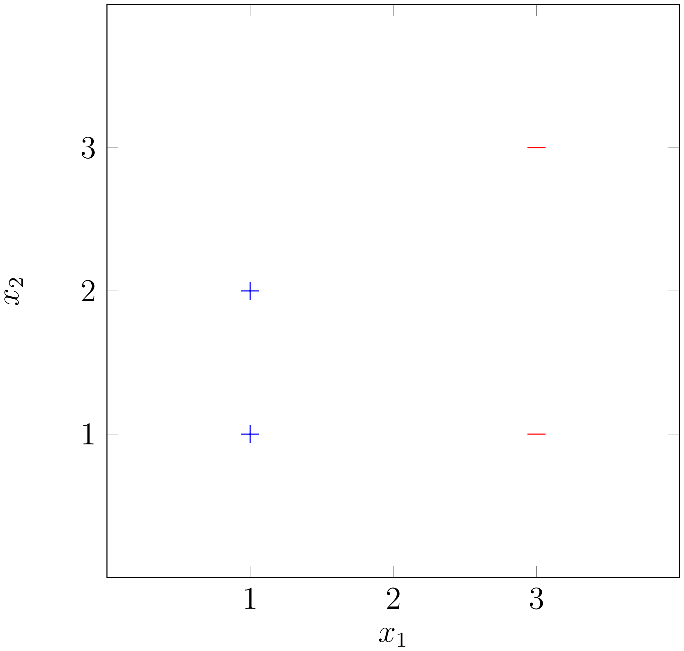
1. One practical issue that arises in real-world classification tasks is that of imbalanced data, i.e., when one class is far more prevalent than all other classes. Consider, for example, a binary classification task with $y \in \{-1, +1\}$ where 99% of all possible data points have label -1: a simple classifier that always predicts -1 will achieve an accuracy of 99%. This can make it difficult to improve on/assess whether a particular machine learning model has learned anything at all! One way of learning in this setting is to penalize errors on the rare class more than errors on the common class.

   (a) **Math:** You decide to apply SVMs to the setting described above. Formulate a *primal* optimization problem that defines a soft-margin SVM where the soft error on training data points with label $+1$ is penalized 50 times more than the soft error on training data points with label -1 (**Hint:** the constraints associated with your optimization problem should be the same as the one presented in lecture, only the objective function will differ).

(b) **Math:** Write out the Lagrangian function and its partial derivatives w.r.t. the optimization variables.

(c) **Math:** Write down the dual of the primal optimization problem in this setting.

2. Consider the following dataset, consisting of two data points with label + and two data points with label −



(a) **Drawing:** A non-essential support vector is a support vector that if removed, would not result in a different decision boundary. Conversely, an essential support vector is a support vector that if removed, would change the decision boundary.

On the figure above, draw the decision boundary you would learn using a hard-margin SVM on this dataset and circle all of the *non-essential* support vectors.

(b) **Drawing:** Again on the figure provided above, draw a new data point with the label + such that

- the hard-margin SVM decision boundary changes,
- the number of essential support vectors *increases* and
- the number of non-essential support vectors *decreases*.

8 Kernels

1. Car-talk statistician Marge Innovera proposes the following simple kernel function:

$$K(x, x') = \begin{cases} 1 \text{ if } x = x' \\ 0 \text{ otherwise.} \end{cases}$$

(a) **Math:** Prove this is a legal kernel. You may assume the feature space $\mathcal{X}$ is finite. Specifically, describe an implicit mapping $\Phi : \mathcal{X} \to R^m$ (for some value m) such that $K(x, x') = \Phi(x) \cdot \Phi(x')$.

(b) **True or False:** In the Φ-space, any labeling of the points in $\mathcal{X}$ will be linearly separable so we can always run a kernelized version the hard-margin SVM. Briefly justify your answer in 2-3 concise sentences.

○ True

○ False

2. **Short answer:** In 2-3 concise sentences, briefly describe the primary benefit of the kernel trick.

3. **Select all that apply:** If k is a valid kernel, which of the following statements must be true?

   □ $k(x, x') = k(x', x)\ \forall\ x$ and x'

   □ $k(x, x') \geq 0\ \forall\ x$ and x'

   □ If $k(x, x) = k(x', x')$, then $x = x'$

   □ $k(x, x') = k(x - x', 0)\ \forall\ x$ and x'

   □ None of the above