10-701 Machine Learning
Fall 2023
Final Practice Problems

1 CNN

1. **Select one:** What is the *primary* purpose of a convolutional layer in a CNN?

   ◯ To reduce the dimensions of the input image.

   ◯ To detect features such as edges and textures in the input image.

   ◯ To classify the input image into various categories.

   ◯ To flatten the input image across channels.

2. Suppose you are building a CNN that takes input images with 3 channels, each of size 20×20 pixels. Your first convolutional layer takes the 3 input channels and produces 5 output channels; it uses 4×4 filters. It also uses a padding of 2 (along all sides of the image) and a stride of 2 (in both dimensions).

   i. **Numerical answer:** How many parameters are there in this convolutional layer, including the bias terms?

   ii. **Numerical answer:** What is the dimensionality of the output of this convolutional layer? Include the channel dimension in your answer and make sure you clearly indicate which dimension is the channel dimension.

3. **Select one:** In CNNs, what is the *primary* advantage of using an inception block over a conventional convolutional layer?

   ◯ It allows for simultaneous use of filters of different sizes to capture features at various scales.

   ◯ It simplifies the network architecture by reducing the number of layers needed to learn the same set of features.

   ◯ It reduces the total number of parameters, making the network easier to train.

   ◯ It enhances the non-linearity of the network without increasing the depth.

2 Unsupervised Learning

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

   i. 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

   ii. The k-means algorithm will always converge to the globally optimal solution.

      ◯ True

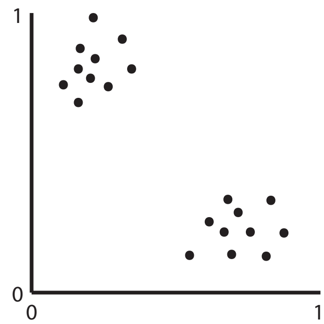
      ◯ False

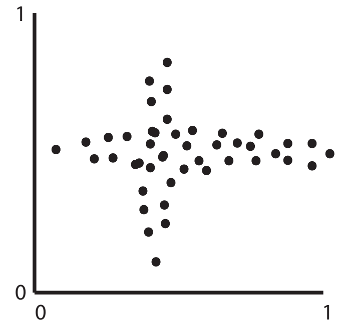
   iii. In the k-means algorithm, the objective function's value can increase, decrease or stay the same after each iteration.

      ◯ True

      ◯ False

2. 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.





3. 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.

   i. **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?

      ○ $E < E'$

      ○ $E \leq E'$

      ○ $E = E'$

      ○ $E \geq E'$

      ○ $E > E'$

   ii. **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?

      ○ $E < E'$

      ○ $E \leq E'$

      ○ $E = E'$

      ○ $E \geq E'$

      ○ $E > E'$

4. **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.

3 VAEs & GANs

1. **Math:** Suppose you have a black box which generates values from a normal distribution with mean 1 and variance 4 i.e., $x_i \sim \mathcal{N}(\mu = 1,\ \sigma^2 = 4)$. Using the *reparametrization trick* for VAEs, write down a formula for $y_i \sim \mathcal{N}(\mu = -1,\ \sigma^2 = 9)$ in terms of x_i.

2. **Short answer:** Recall that the ELBO objective function used to train VAEs is short for "Evidence Lower BOund". In 2-3 concise sentences, briefly describe the quantity that the ELBO lower bounds and explain why it is used as the objective function.

3. **Select one:** Which of the following best describes the relationship between the discriminator and the generator in a GAN?

   ◯ Given an unlabelled dataset, the generator generates labels for the data points and the discriminator computes the likelihood of the generated labels.

   ◯ Given an unlabelled dataset, the discriminator clusters the data points and the generator generates new data points, conditioned on cluster assignments.

   ◯ Given an unlabelled dataset, the generator generates latent representations for each data point and the discriminator predicts whether some input vector is the latent representation for some data point from the dataset.

   ◯ Given an unlabelled dataset, the generator generates data points and the discriminator classifies data points as coming from the dataset of the generator.

4. **Numerical answer:** Suppose that the probability of a some data point x is $p_{data}(x) = \frac{1}{10}$ and the probability of x under a generator G is $p_G(x) = \frac{1}{2}$. If D is the optimal discriminator in this setting, what is the probability that D assigns to x of being from the generator?

5. **Short answer:** In a BEGAN, the binary discriminator is replaced with an autoencoder. In 1-2 concise sentences, describe the objective function that a BEGAN uses to train the discriminator.

4 RNN

1. **Select all that apply:** Which of the following statements is/are correct?

   - ☐ Training RNNs is difficult because of vanishing and/or exploding gradients.
   - ☐ Gradient clipping is an effective technique to address the vanishing gradient problem.
   - ☐ RNNs differ from feed-forward neural networks in that they have an additional weight matrix connecting hidden layers across time-steps.
   - ☐ RNNs can process sequences of arbitrary length, while feed-forward neural networks can not.
   - ☐ None of the above.

2. **Select one:** How do Long Short-Term Memory (LSTM) units address some of the limitations of standard RNNs?

   - ◯ LSTMs have lower computational costs, allowing for faster processing of sequences.
   - ◯ LSTMs simplify the structure of RNNs by decreasing the number of parameters, making them easier to train and less prone to overfitting.
   - ◯ LSTMs improve the feature extraction capabilities of RNNs.
   - ◯ LSTMs address the vanishing gradient problem through memory cells.

3. **Short answer:** In 2-3 concise sentences, briefly describe the bi-directional RNN architecture and describe why they are *not* appropriate for language modelling.

5 Attention & Transformers

1. **Select all that apply:** For a fixed input-size and embedding dimension, which of the following statements is *not* true about multi-head attention relative to single-head attention?

   ◯ Multi-head attention is more suitable for parallel computation than single-head attention.

   ◯ Multi-head attention layers have more total parameters in their query, key and value matrices than single-head attention layers.

   ◯ Multi-head attention is compatible with more attention score functions than single-head attention.

   ◯ Multi-head attention are able to capture more diverse relationships between tokens than single-head attention.

   ◯ None of the above

2. Suppose you have a transformer model that employs multi-head attention. The inputs to your model are sequences of T tokens and each token is represented by a d_M-dimensional embedding. Your model has H heads and for each head, the dimensionality of the key and query vectors is d_K and the dimensionality of the output vectors is d_V.

   i. **Math:** What are the dimensions of the key matrix *for one of the attention heads* i.e., K^h where $K^h = XW^h$?

   ii. **Math:** What is the dimensionality of the multi-headed attention output *before any sort of concatenation*?

3. **Short answer:** In 2-3 concise sentences, briefly explain why positional encodings are used in transformer models.

6 Pre-Training, Fine-Tuning and In-Context Learning

1. **True or False:** A model finetuned on a masked language modelling objective is appropriate for generative tasks such as text completion. Briefly justify your answer in 1-2 concise sentences

   ◯ True

   ◯ False

2. **Short answer:** Give an example of chain-of-thought prompting. For full credit, your response must include the question you wish to be answered and cannot be one of the examples provided in lecture.

3. **Select all that apply:** Which of the following techniques directly update the parameters of a pretrained language model?

   ☐ Reinforcement learning from human feedback (RLHF)

   ☐ Instruction finetuning

   ☐ Few-shot learning

   ☐ Soft-prompting

   ☐ None of the above

7 Robustness

1. **Select one:** What is the primary difference between targeted and untargeted adversarial attacks?

   ◯ Targeted attacks attempt to manipulate a model towards a specific class whereas untargeted attacks solely care about increasing model loss.

   ◯ Targeted attacks limit the magnitude of allowable perturbations whereas untargeted attacks allow for arbitrary perturbations.

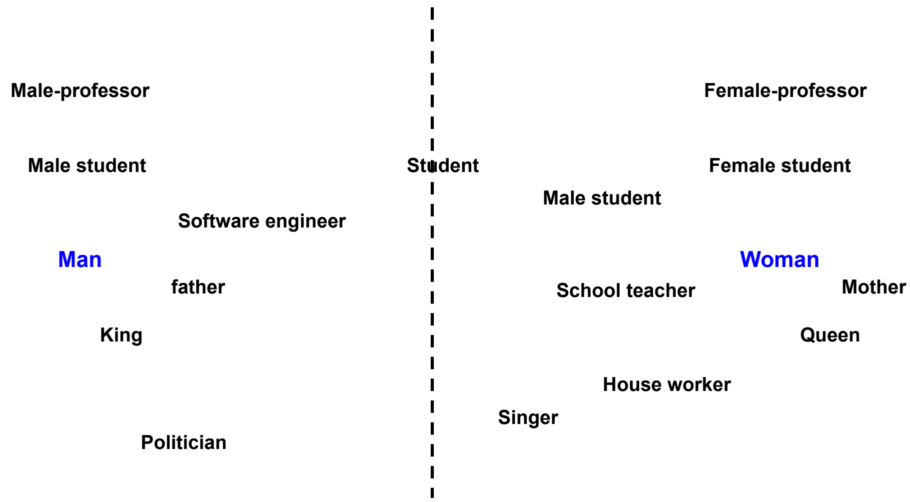
   ◯ Targeted attacks can be optimized via gradient-based methods because the objective is differentiable whereas untargeted attacks cannot.

   ◯ Targeted attacks can be performed in multiclass classification settings whereas untargeted attacks can only be performed in binary classification settings.

2. **Short answer:** In 2-3 concise sentences, briefly describe the primary difference between label shift and covariate shift; for full credit, your answer must address the assumptions made in both settings in terms of conditional probabilities.

3. **Proof:** Given a fixed classifier f, prove that the expected, column-normalized confusion matrices of f under two different label distributions p and q will be identical.

8 Bias



1. Consider the pre-trained word embedding space shown in the figure above.

   i. **Select all the apply:** Which of the following is a potential indicator that this word embedding is gender-biased in terms of occupation?

   - ☐ ”man” - ”king” = ”woman” - ”queen”
   - ☐ ”man” - ”software engineer” = ”woman” - ”school teacher”
   - ☐ ”male professor” - ”male student” = ”female professor” - ”female student”
   - ☐ ”male professor” - ”student” ≠ ”female professor” - ”student”
   - ☐ None of the above

   ii. **Short answer:** Word embeddings like the one shown above are generally trained on web-scale text corpuses. In 1-2 concise sentences, briefly explain how gender bias, as defined by the spatial relationship between pairs of non-gendered words, is learned by word embeddings.

2. **Select one:** What is the relationship between demographic parity, separation and calibration in the general case?

   ◯ All three of these conditions can hold simultaneously.

   ◯ Any one of these conditions is not achievable if either of the other conditions is true.

   ◯ Any one of these conditions is not achievable if both of the other conditions are true.

   ◯ None of these conditions are achievable in the general case.

9 Interpretability

1. **Short answer:** In 2-3 concise sentences, define an integrated gradient for a deep learning model and identify the primary issue with using an integrated gradient for explaining a model's predictions.

2. **Select one:** Which of the following is the best description of a counterfactual explanation?

   ◯ Given an input that was predicted to be of one class, find the single feature that needs to be changed the least in order to change the model's prediction.

   ◯ Given an input that was predicted to be of one class, find the smallest set of features that if fixed, make it so that the model's prediction can never change regardless of the other features' values.

   ◯ Given an input that was predicted to be of one class, find the radius of the smallest ball centered at the input that includes a data point of another class.

   ◯ Given an input that was predicted to be of one class, find the nearest data point that would have been predicted to be of another class.