

EECS545 Lecture 2 Quiz Solutions

January 14, 2024

Question 1. $y = \frac{c+1}{3}$.

Let $y = \alpha$. Then, the total least squared loss is $E = \alpha^2 + (c - \alpha)^2 + (1 - \alpha)^2 = 3\alpha^2 - (2c + 2)\alpha + (c^2 + 1)$. Minimizing for α , we get $\alpha = \frac{c+1}{3}$.

Question 2. $\frac{c^2+5}{2}$

Plugging in values:

$$\Phi w = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix}$$

Then, the loss (in vectorized form) is

$$E(w) = \frac{1}{2} \|\Phi w - y\|^2 = \frac{1}{2} \left\| \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix} - \begin{bmatrix} 0 \\ c \\ 1 \end{bmatrix} \right\|^2 = \frac{1}{2} \left\| \begin{bmatrix} -1 \\ -c \\ -2 \end{bmatrix} \right\|^2 = \frac{c^2 + 5}{2}$$

Question 3. $\begin{bmatrix} c+3 \\ 1 \\ 2 \end{bmatrix}$

Plugging in:

$$\nabla E(w) = \Phi^\top (\Phi w - y) = \begin{bmatrix} 1 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ -c \\ -2 \end{bmatrix} = \begin{bmatrix} -c-3 \\ -1 \\ -3 \end{bmatrix}$$

Gradient update:

$$w' = w - \eta \nabla E(w) = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix} - \begin{bmatrix} -c-3 \\ -1 \\ -3 \end{bmatrix} = \begin{bmatrix} c+3 \\ 1 \\ 2 \end{bmatrix}$$

Question 4. True.

Conceptually, we are fitting a quadratic polynomial to three points. We know from algebra there is always an optimal solution that fits the points perfectly.

Question 5. False.

Conceptually, we are fitting a linear equation to three points. However, if $c \neq \frac{1}{2}$, then there is no way to fit this equation without some error.

EECS545 Lecture 3 Quiz Solutions

1. Let $D = \{(x_i, y_i)\}_{i=1,\dots,3} = \{(-1, -0.5), (0, 0), (1, 1)\}$. Suppose that we want to fit a linear regression model with a 1-degree polynomial function with $\hat{y} = w_0 + w_1x$. After convergence, what would be the solution of w_0 and w_1 ? Hint: write down the objective function and set the derivative with respect to w_0 and w_1 to zero to find the optimal solution.

Solution: Plugging in for the objective function:

$$E(w) = \frac{1}{2} [(w_0 - w_1 + 0.5)^2 + (w_0)^2 + (w_0 + w_1 - 1)^2]$$

Partial derivatives:

$$\frac{\partial E(w)}{\partial w_0} = (w_0 - w_1 + 0.5) + w_0 + (w_0 + w_1 - 1) = 3w_0 - 0.5$$

$$\frac{\partial E(w)}{\partial w_1} = -(w_0 - w_1 + 0.5) + (w_0 + w_1 - 1) = 2w_1 - 1.5$$

Setting these to zero, we get $w_0 = \frac{1}{6}, w_1 = \frac{3}{4}$.

2. Continued. Suppose instead we used ridge linear regression with $\lambda = 1$. Find w_0 and w_1 after convergence.

Solution: Plugging in for the objective function:

$$E(w) = \frac{1}{2} [(w_0 - w_1 + 0.5)^2 + (w_0)^2 + (w_0 + w_1 - 1)^2] + \frac{1}{2}(w_0^2 + w_1^2)$$

Partial derivatives:

$$\frac{\partial E(w)}{\partial w_0} = (w_0 - w_1 + 0.5) + w_0 + (w_0 + w_1 - 1) + w_0 = 4w_0 - 0.5$$

$$\frac{\partial E(w)}{\partial w_1} = -(w_0 - w_1 + 0.5) + (w_0 + w_1 - 1) + w_1 = 3w_1 - 1.5$$

Setting these to zero, we get $w_0 = \frac{1}{8}, w_1 = \frac{1}{2}$.

3. Consider polynomial regression by optimizing the least-squares objective function with regularization term λ . **Choose all options that apply:**
- (a) As λ increases, the curve formed by polynomial regression will become flatter.
 - (b) Best practice indicates that we should keep tuning λ until we find a final value λ^* that minimizes error on the test set.

- (c) It is not valid to use the $L1$ norm in place of the $L2$ norm in the regularized objective function.

Solution:

- (a) True.
- (b) False.
- (c) False.

EECS545 Lecture 4 Quiz Solutions

1. Which of the following is a disadvantage of Newton's method compared to gradient descent?:
 - (a) Newton's method usually takes more iterations than gradient descent to converge.
 - (b) Newton's method takes more time to compute on an individual iteration than gradient descent.
 - (c) Newton's method does not always find the optimum for a convex function.
 - (d) Newton's method requires more training data to apply, compared to the gradient descent.

Solution: (b). Need to perform matrix inversion (which is expensive) in Newton's method but not in gradient descent.

2. Suppose we run one iteration of Newton's method on $f(x) = x^3 - 2x^2 + 4$ (in the interval $[0, 5]$). If $x_0 = 2$, what will be the first approximation x_1 ? Hint: start from computing the first derivative. (note: any answer within the error bound of 0.01 will be marked as correct.)

Solution: $f'(x) = 3x^2 - 4x$ $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} = 2 - \frac{4}{4} = 1$

3. Continued from Q2. What would be the second approximation x_2 ?

Solution: $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)} = 1 - \frac{3}{-1} = 4$

4. (True/False) Logistic regression learns a non-linear decision boundary because the logistic function is non-linear.

Solution: False. Logistic regression learns a linear decision boundary in the feature space. In other words, when you use x as features, then the decision boundary is a linear function of x . If you use non-linear function $\phi(x)$ as features, the decision boundary can be non-linear function of x , but this is not because the logistic function is non-linear.

EECS545 Lecture 5 Quiz Solutions

1. Suppose for a balanced dataset for binary classification (e.g., positive and negative examples are approximately 50% each for both train and test set) with n examples we run classification using K-nearest neighbors, with $K = n$. For simplicity, we assume that K is an odd number to make tie-breaking unambiguous. In this case, K-nearest neighbors is... :
 - (a) The accuracy of this model would be higher than 90% over the test set.
 - (b) n is a good choice for K .
 - (c) $K = n$ leads to a high bias.
 - (d) Leads to a classifier that ignores the input.

Solution: (c) and (d). (a) The accuracy of this model would be approximately 50%, so (b) n is not a good choice for K .

2. Let's say we have four training examples $\{(x_1, x_2), y\} = [\{(-3, 3), 0\}, \{(4, 0), 1\}, \{(3, 4), 0\}, \{(0, -6), 1\}]$. When we provide $(0, 0)$ as the query example and set $K=3$ with L^2 norm as the distance function, what will be the predicted output class?

Solution: 0. L^2 distances from $(0,0)$ are $3\sqrt{2}(\approx 4.24), 4, 5, 6$. Since the model will not choose $\{(0, -6), 1\}$, the majority class among three points will be 0.

3. Continued from Q3. If we change the distance function from L^2 norm to L^1 norm, what will be the predicted output class?

Solution: 1. L^1 distances from $(0,0)$ are 6, 4, 7, 6. Because $\{(3,4), 0\}$ will be excluded, the majority class will be 1.

4. Select all that true:
 - (a) Generative models in general learn fewer parameters than discriminative models.
 - (b) Generative models model the joint probability distribution $p(x, C)$, where x is data and C is class.
 - (c) Generative models that models the joint distribution of data x and classes C can be converted to calculate $p(C | x)$
 - (d) Regularization can only be used for discriminative models and not generative models

Solution: (b) and (c).

5. Suppose we are using GDA on a dataset with two classes. What happens when we use different covariance matrices for each class? Select all that are true:

- (a) Different covariance allows us to model a non-linear decision boundary
- (b) Calculating the MLE for different covariance takes less computation
- (c) Different covariance can increase the log likelihood
- (d) Different covariance is less likely to underfit the training data
- (e) Different covariance guarantees lower test error

Solution: (a), (c), and (d).

Learning different covariance in GDA is analogous to adding more polynomial features to linear regression.

Suppose the training data is non-linear or skewed, and imagine we fit (i) GDA with fixed covariance and (ii) GDA with learned covariance. Then the likelihood for (ii) should always be higher because (ii) subsumes (i).

When the training data is skewed (e.g. oval shaped) learning different covariance can learn this skew better than when the covariance is fixed, hence the log likelihood increases and it is less likely to underfit.

EECS545 Lecture 6 Quiz Solutions

1. **Select all that are true.**

- (a) Consider a problem where you want to use a high-dimensional features (where there may be some correlation between the features). Between Naive Bayes and Logistic Regression, Naive Bayes is the better choice.
- (b) Naive Bayes classifier and GDA (Gaussian Discriminant Analysis) are generative models.
- (c) Laplacian smoothing for Naive Bayes avoids zero product for words that show up as only spam / only non-spam

Solution: (b) and (c).

(a) is not true: naive Bayes assumes conditional independence of features given class labels. This may be a too strong assumption when there is non-trivial correlation between features.

2. Naive Bayes practice. Consider the following dataset $\{(\text{spam or not spam}, [\text{tokens}])\} = \{(\text{spam}, [\text{A}, \text{B}, \text{B}, \text{A}]), (\text{not spam}, [\text{C}, \text{A}, \text{B}]), (\text{not spam}, [\text{B}, \text{A}, \text{B}])\}$. How many words (vocabulary size M in the lecture) exist in this dataset?

Solution: $M = 3$ (A, B, C).

3. Continued. Find the naive bayes MLE estimate for $P((\text{spam}, [\text{C}, \text{A}, \text{B}, \text{B}, \text{A}]))$ without laplacian smoothing.

Solution: $\mu_C^{\text{spam}} = 0$, so the entire likelihood is 0.

4. Continued. Find the MLE estimate for $P((\text{spam}, [\text{C}, \text{A}, \text{B}, \text{B}, \text{A}]))$ with laplacian smoothing. We still assume that each token t_i is independent.

Solution:

$$\phi^{spam} = \frac{1}{3} \quad (1)$$

$$\mu_A^{spam} = \frac{2+1}{4+3} = \frac{3}{7} \quad (2)$$

$$\mu_B^{spam} = \frac{2+1}{4+3} = \frac{3}{7} \quad (3)$$

$$\mu_C^{spam} = \frac{0+1}{4+3} = \frac{1}{7} \quad (4)$$

$$P((spam, [C, A, B, B, A]) = \mu_C^s (\mu_A^s)^2 (\mu_B^s)^2 \phi^{spam} = \frac{27}{16807} \quad (5)$$

EECS545 Lecture 7 Quiz Solutions

1. Overfitting is characterized by:
 - (a) Low variance and low bias
 - (b) Low variance and high bias
 - (c) High variance and low bias
 - (d) High variance and high bias

Solution: (c). Overfitting is the case when there is a high variance and low bias.

2. Consider polynomial regression by optimizing the least-squares objective function with regularization: $\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^\top \phi(x^{(i)}) - y^{(i)})^2 + \lambda \|\mathbf{w}\|^2$ where $x^{(i)}$ is scalar and $\phi(x^{(i)})_j = (x^{(i)})^j$. **Choose all options that apply:**
 - (a) It is not valid to use the $L1$ norm in place of the $L2$ norm in the regularized objective function.
 - (b) Increasing λ increases estimator bias and decreases estimator variance.
 - (c) Best practice indicates that we should keep tuning λ until we find a final value λ^* that minimizes error on the test set.
 - (d) As λ increases, the curve formed by polynomial regression will become flatter.

Solution: (b) and (d). (a) You can choose L1 if it is better suited for the problem you are trying to solve. (c) You must not use the test set to determine the value of a hyper-parameter.

3. (True/False) Training with more training data usually improves generalization. Suppose we have a training set containing N training cases. Suppose we duplicate the training cases L times each so that the training set now contains $L \times N$ cases (but still only N unique cases). Will this improve the generalization performance of a logistic regression (in any case, assume that the regularization (e.g., L_2) hyperparameter will be carefully tuned on the validation data)?

Solution: False. Naively duplicating data does not help since the objective function will be increased by a constant factor. So, the optimal solution would remain the same as in the case of using original data.

4. (True/False) Suppose you are using logistic regression (trained with maximum likelihood), and the classifier is performing poorly (has an unacceptably high error) on the test set but performing well on the training set. In this case, adding more training examples will likely increase the training error. (For simplicity, we assume that both the training and testing examples were randomly sampled from an IID distribution.)

Solution: True. because it increases the difficulty of fitting (compared to the original setting)

5. (True/False) (Continued from Q4) Adding more training examples will likely increase the test error.

Solution: False. The model has seen more training data and may cover the data space more densely and thus generalize better.

EECS545 Lecture 8 Quiz Solutions

1. Which of the following are true statements about kernels? **Choose all options that apply:**
- (a) A machine learning algorithm can be kernelized if it does not need explicit access to the feature vectors and instead only requires access to inner products of the feature vectors.
 - (b) A Gram/kernel matrix must be positive semidefinite.
 - (c) The product of two kernel functions is still a kernel function.
 - (d) The sum of two kernel functions is always a kernel function.

Solution: (a), (b), (c), (d). All of them are true.

2. What is the purpose of the kernel trick?
- (a) To transform the problem from regression to classification.
 - (b) To transform the data into a richer feature space without explicitly computing the feature vector.
 - (c) To transform the problem from supervised to unsupervised learning.
 - (d) To transform a linear regression model to SVM.

Solution: (b) Please check the slide 10 of Lecture 8 for an example.

3. (True/False) For any two documents \mathbf{x} and \mathbf{z} , define $k(\mathbf{x}, \mathbf{z})$ to equal the number of unique words that occur in both \mathbf{x} and \mathbf{z} (i.e., the size of the intersection of the sets of words in the two documents \mathbf{x} and \mathbf{z}). This function cannot be considered as a kernel.

Solution: False. We can make k as kernel by setting $\phi(\mathbf{x})$ as a binary vector whose i -th entry is 1 when the document \mathbf{x} contains the i -th word and 0 if it doesn't.

4. (True/False) $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) - k_2(\mathbf{x}, \mathbf{y})$ is a kernel if k_1 and k_2 are valid kernels, and those kernels are defined as $k_1(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \diamond \mathbf{y} + 1)^2$ and $k_2(\mathbf{x}, \mathbf{y}) = \mathbf{x} \diamond \mathbf{y}$ (Please assume \diamond is a predefined operation that can lead k_1 and k_2 valid kernel. For example, $\diamond(\mathbf{x}, \mathbf{y}) = 2\mathbf{x} + \frac{\mathbf{y}}{2}$ could work as a candidate).

Solution: True. $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \diamond \mathbf{y})^2 + \mathbf{x} \diamond \mathbf{y} + 1$ with each term in the sum being a kernel.

5. (True/False) A Gaussian kernel will always have better classification performance at a testing time compared to a linear kernel.

Solution: False. Bad hyperparameter choice or small (finite) training data size setting could lead the kernel to overfit.

EECS545 Lecture 9 Quiz Solutions

1. Which of the following can only be used when training data are linearly separable? **Choose all options that apply:**

- (a) Linear hard-margin SVM.
- (b) Linear Logistic Regression.
- (c) Linear Soft-margin SVM.

Solution: (a). Please revisit Slide 12.

2. Choose all true statements regarding soft-margin SVM:

- (a) When the regularization parameter C increases in soft-margin SVM, the number of support vectors will increase.
- (b) When the regularization parameter C increases in soft-margin SVM, the total sum of slack variables will decrease.
- (c) For a fixed linearly-separable dataset, the minimum objective value for soft-margin SVM is at least as small as the minimum objective value for hard-margin SVM.
- (d) When the regularization parameter C increases in soft-margin SVM, the margin of the resulting classifier will increase.

Solution: (b), (c).

(a), (b), (d): Increasing C would penalize more for the slack variables, decreasing the total sum of slack variables at optimum. So, it will not increase the margin. The number of support vectors is not directly correlated.

(c): Since the solution for the hard-margin SVM (0 slack variables) is still a feasible point for the soft-margin SVM problem, having more freedom on slack variable can only decrease the minimum for the soft-margin SVM.

3. (True/False) The farthest examples from the decision boundary of a dataset are called “support vectors”. (Assume the dataset has extremely large number of examples.)

Solution: False. Examples farthest from the decision boundary will not be the support vector as it will not contribute in determining the margin. Please check the plot in Slide 10.

4. (True/False) After training a linear SVM classifier, you observe that the test error is high while the training error is low. Increasing the value of parameter C is likely to help improve test performance.

Solution: False. Decreasing C will improve margin and that might help.

5. (True/False) Consider a supervised learning problem in which the training examples are points in 2-dimensional space. The positive examples are $(1, 1)$ and $(-1, -1)$. The negative examples are $(1, -1)$ and $(-1, 1)$. Note that the positive examples are NOT linearly separable from the negative examples in the original space. Feature transformation $\phi(\mathbf{x}) = [1, x_1, x_2, x_1x_2]$, where x_1 and x_2 are the first and the second coordinates of a generic example \mathbf{x} , could help the SVM to separate the positive and negative examples.

Solution: True. $\mathbf{w} = [0, 0, 0, 1]$ satisfies the maximum-margin decision surface separating the positive and negative examples on the feature space.

EECS545 Lecture 10 Quiz Solutions

1. It is given that the constrained optimization problem below is “convex”. Which of the following statements are correct? (**Choose all options that apply**)

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, i = 1, \dots, m \\ & h_i(\mathbf{x}) = 0, i = 1, \dots, p \end{aligned}$$

- (a) The objective function f is convex.
- (b) All g_i are convex.
- (c) All h_i are affine.
- (d) All h_i are convex but not necessarily affine.

Solution: (a), (b), (c) Please revisit Slide 16.

2. Which of the following is always true about \star ? Note that \mathcal{L} is the Lagrangian function corresponding to the optimization problem.

$$\star = \min_{\mathbf{x}} \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0, \forall i} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) - \max_{\boldsymbol{\nu}, \boldsymbol{\lambda}: \lambda_i \geq 0, \forall i} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

- (a) $\star = 0$
- (b) $\star \geq 0$
- (c) $\star \leq 0$
- (d) $\star > 0$
- (e) $\star < 0$

Solution: (b). Primal optimization (the first term) is greater or equal to Dual optimization (the second term); weak duality. Those two terms are the same if conditions are met (= strong duality). See Slide 13-14 for details.

3. (True/False) Linear hard-margin SVM must have at least one support vector for positive margin and one support vector for negative margin that satisfies the margin constraint. (Let's assume we have trained linear hard-margin SVM with linearly separable data, for simplicity)

Solution: True. If there is no support vector in either end, it means there is a linear hard-margin SVM that have a larger margin.

4. Let $(x^{(n)}, y^{(n)})$ be a data point that is a support vector in the Linear Soft SVM formulation. Which of the following equations does the support vector satisfy?

(a) $y^{(n)} h(\mathbf{x}^{(n)}) = 1$

(b) $y^{(n)} h(\mathbf{x}^{(n)}) = 1 - \xi^{(n)}$

Solution: (b). Please see Slide 29. (a) is for the Linear hard-margin SVM case (as in Slide 27).

5. (True/False) The dual view of SVM shows that the objective function depends on $\phi(\mathbf{x})$ only via inner products $(\phi(\mathbf{x})^\top \phi(\mathbf{x}))$ and hence the kernel trick can be used.

Solution: True. See Slide 25-26 for an example.

EECS545 Lecture 11 Quiz Solutions

1. Assume we have a fully-connected neural network with 1 hidden layer with ReLU activations ($h(a) = \max(0, a)$) for binary classification. Which of the following statements are true about the behavior of the network? (**Choose all options that apply**)
- (a) The total training time is always the fastest for the smallest possible batch size since each gradient step takes less time.
 - (b) This model will have a non-linear decision boundary.
 - (c) Adding more layer sometimes perform worse than shallow networks.
 - (d) Multiplying all the weights and biases in the network by a factor of 10 after training the network will not change its classification accuracy.

Solution: (b),(c),(d).

(a): Although each gradient update is faster for a small batch size, the number of updates is larger. Also, the train time varies depending on the dataset and the environment.

(b): True, as ReLU activation brings nonlinearity.

(c): True. Deep networks trained with backpropagation (without any sort of unsupervised pretraining) sometimes perform worse than shallow networks due to overfitting (See Slide 67)

(d): The logit values may change, but the predictions remain the same.

2. For the sigmoid activation function and the ReLU activation function, which of the following are true in general? (**Choose all options that apply**)
- (a) Both activation functions are monotonically non-decreasing
 - (b) Both functions have a monotonic first derivative
 - (c) Compared to the sigmoid, the ReLU is more computationally expensive
 - (d) The first derivative of ReLU is quadratic.

Solution: (a).

(a) True. Simply graph the activation functions

(b) False. Sigmoid has non-monotonic derivative $\sigma(x)(1 - \sigma(x))$

(c) and (d) False. ReLU is simpler as all positives have derivative 1 and all negatives have 0.

3. (True/False) Logistic regression can be viewed as a single-layer neural network (no hidden layer) without any non-linear activation before applying softmax in the output layer.

Solution: True. Both of them will learn $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$.

4. (True/False) When initializing weights in a fully connected Neural Network, we should set the weight to 0 in order to preserve symmetry across all neurons.

Solution: False. We should not set it to 0 to help hidden units do not get the same gradients from the beginning.

5. (True/False) Any multi-layer neural network with linear activation functions for all hidden layers can be represented as a neural network without any hidden layer.

Solution: True. If linear activation functions are used for all the hidden units, output from hidden units will be written as a linear combination of input features.

EECS545 Lecture 12 Quiz Solutions

1. Which of the following are the benefits of using CNNs, as opposed to fully connected networks, for image recognition tasks? (**Choose all options that apply**)
- (a) The ability to express a wider variety of more complicated functions of the input features.
 - (b) Fewer model architecture hyperparameters for the designer to select.
 - (c) Enables the network to more easily learn and recognize features regardless of their position in the image.
 - (d) Typically requires fewer data to train well.

Solution: (c) and (d)

- (a) False. Compared with CNNs, fully connected networks offer greater expressivity and model capacity as they have many more weights.
- (b) False. CNNs involve more architecture parameters (kernel size, stride, dilation, padding, pooling, etc.).
- (c) True. Since CNNs share weights, the same features can be learned and recognized at different positions
- (d) True. Fewer parameters often imply less data is required to train a reasonable model.

2. Which of the following statement are true about batch normalization in neural network training? (**Choose all options that apply**)
- (a) It addresses overfitting issue.
 - (b) It restricts activations from becoming too high or low.
 - (c) It makes training faster.
 - (d) It is a non-linear transformation to center the training dataset around the origin.

Solution: (b) and (c).

- (a) False. Batch norm does not change the number of parameters to learn. (d) False. no non-linear function is introduced in batch normalization.

3. How many parameters do we need to learn for the following network structure? A $32 \times 32 \times 3$ image input, followed by a convolution layer with 8 filters of size 3×3 (stride 1 and 1 pixel of zero-padding), then another convolution layer with 4 filters of size 2×2 (stride 2 and no zero-padding), and finally a max-pooling layer with a 2×2 filter (stride 2 and no zero-padding). (Note: the depth of all filters are not explicitly spelled out, and we assume no bias/intercept terms are used here.)
- (a) 59
 - (b) 88

- (c) 264
- (d) 344

Solution: (d). $8 \times (3 \times 3 \times 3) + 4 \times (2 \times 2 \times 8) = 344$

4. (Continued from 3) What is the final output dimension of the last question?

- (a) $8 \times 8 \times 1$
- (b) $16 \times 16 \times 1$
- (c) $8 \times 8 \times 4$
- (d) $16 \times 16 \times 4$

Solution: (c). $32 \times 32 \times 3$ -(1st conv)- $32 \times 32 \times 8$ -(2nd conv)- $16 \times 16 \times 4$ -(maxpool)- $8 \times 8 \times 4$

5. (True/False) Using Momentum, instead of SGD, always helps the network converge faster.

Solution: False. Use of active learning (including Momentum) is reasonable in training a network, but too much momentum will be harmful in training a network.

EECS545 Lecture 13 Quiz Solutions

March 7, 2024

1. What are the benefits of using an RNN over applying CNN over time (i.e., 1D CNN)? (**Choose all options that apply**)

- (a) RNNs require less inference time than CNNs in Language modeling.
- (b) RNNs can better understand the sequential dependencies.
- (c) RNNs can better handle sequences with unknown lengths.
- (d) Training RNNs are easier than CNNs as RNNs are less likely to have gradients explode/vanish issue.

Solution: (b) and (c)

(a): RNNs need to infer each word one by one during inference time (See Slide 69).
(d): RNNs are more likely to suffer gradient explode/vanish issues.

2. Which tricks can help address the exploding/vanishing gradient problem? (**Choose all options that apply**)

- (a) Use LSTM instead of vanilla RNN.
- (b) Use sigmoid activation instead of tanh in vanilla RNN.
- (c) Use orthogonal initialization.
- (d) Use gradient clipping.

Solution: (a) (c) (d)

Please revisit the slides from page 38). (b) would still suffer exploding/vanishing gradient issues; tanh can be written as $2\sigma(2x) - 1$ if sigmoid is $\sigma(x)$.

3. (True/False) The vanishing gradient in RNN could cause the parameters to be biased to capture short- term dependencies.

Solution: True. Long-term dependencies in the sequence become affected as the gradient vanishes in the middle (Slide 42)

EECS545 Lecture 14 Quiz Solutions

1. What is the main function of the attention mechanism in deep learning?
 - (a) To reduce overfitting in the neural network
 - (b) To focus on the most important parts of the input sequence
 - (c) To speed up the training of the neural network
 - (d) To regularize the weights in the neural network

Solution: (b).

2. What are the benefits of using a attention over an recurrent layers for sequential tasks?
 - (a) Attention is less memory intensive.
 - (b) Attention is more parallelizable during training
 - (c) Attention is more parallelizable for generating a new sequence during test time
 - (d) Attention doesn't suffer from vanishing gradients
 - (e) Attention can directly use inputs from long sequences without memory

Solution: (b) and (e).

In attention during training (b), we can encode the entire sequence length in one pass, whereas RNN requires T passes, where T is the sequence length. However, when inferring a new sequence (c) of some length T , both RNNs and attention have to do T passes.

3. In a transformer model, what is the purpose of the positional encoding?
 - (a) To add information about the position of each token in the input sequence
 - (b) To reduce the dimensionality of the input sequence
 - (c) To compute the attention weights between each pair of tokens in the input sequence
 - (d) To aggregate the information from each token in the input sequence into a single vector

Solution: (a).

4. Path length is the number of layers an input has to go through, before it is output. An RNN has a maximum path length of $O(n)$, where n is the sequence length. This is because the first input goes through n RNN calls then “exits” through the last sequence output. What is the maximum path length of a self-attention layer?

- (a) $O(1)$
- (b) $O(n)$
- (c) $O(n^2)$
- (d) $O(n^3)$

Solution: $O(1)$. Each input in the sequence goes through the same key-query-value matrix product once in self-attention.

5. In a transformer model, why is it important to use causal attention masks in the decoder?

- (a) To prevent the model from overfitting to the training data.
- (b) To allow the model to attend to all positions in the input sequence.
- (c) To avoid introducing future information into the decoding process.
- (d) To increase the model's capacity to handle long input sequences.

Solution: (c). For an input x_t , A causal attention mask masks out all values from x_{t+1}, x_{t+2}, \dots . So, when encoding x_t , the transformer cannot use any information from the “future”.

EECS545 Lecture 15 Quiz Solutions

1. (T/F) The objective function of K-means decreases monotonically.

Solution: True.

2. Which of the following about EM algorithm is true? Select all that apply.

- (a) The EM algorithm can be used for MLE (maximum likelihood) estimate problems involving latent variables.
- (b) The EM algorithm monotonically increases the lower bound of the log-likelihood $L(q, \theta)$.
- (c) If the posterior $P(\mathbf{Z} \mid \mathbf{X}; \theta)$ is tractable, the EM algorithm always monotonically increases the observed data log-likelihood of the data.
- (d) For some complex models where $P(\mathbf{Z} \mid \mathbf{X}; \theta)$ is not tractable, EM monotonically increases the log-likelihood of data.
- (e) The EM algorithm can find the global maximum data likelihood if ran sufficiently long.

Solution: (a), (b), and (c). See the lecture notes.

3. Which of the following is true about the E-step of the EM algorithm? Select all that apply.

- (a) E-step computes complete data log-likelihood
- (b) E-step computes the posterior probability of the latent variables
- (c) E-step updates the parameters of the model
- (d) In a single E-step, the log-likelihood of the observed data is increased.
- (e) In a single E-step, the lower bound $L(q, \theta)$ of the log-likelihood of the observed data is increased.

Solution: (b) and (e). In E-step, we compute the posterior $P(\mathbf{Z} \mid \mathbf{X}; \theta)$ and set it as $q(\mathbf{Z})$ given fixed parameters θ of the model. This increases the lower bound on the log-likelihood of the observed data, but the log-likelihood of the observed data is kept constant because the parameter does not change.

4. In the latent variable models we discussed in the class, each of the following terms means:

- complete likelihood:
- posterior:
- observed data likelihood:

Solution:

- complete likelihood: $p(\mathbf{X}, \mathbf{Z} \mid \theta)$.
- posterior: $p(\mathbf{Z} \mid \mathbf{X}, \theta)$.
- observed data likelihood: $p(\mathbf{X} \mid \theta)$.

Note that

- $p(\mathbf{X} \mid \mathbf{Z}, \theta)$ is called the conditional data likelihood.

EECS545 Lecture 16 Quiz Solutions

1. What is the main goal of PCA?

- (a) To cluster the dataset in a latent subspace
- (b) To reduce the dimensionality of a dataset
- (c) To maximize the correlation between the features in a dataset
- (d) To minimize the variance in a dataset

Solution: (b). PCA is a dimensionality reduction method.

2. Which of the following are true about PCA? Choose all that apply.

- (a) The principal component vectors discovered by PCA are always orthogonal to each other.
- (b) It is possible to kernelize PCA algorithm.
- (c) PCA can be used for feature selection.
- (d) PCA is an supervised learning algorithm.
- (e) PCA requires an assumption that the data is normally (Gaussian) distributed.

Solution: (a), (b), and (c).

(a) The principal component vectors are eigenvectors, which are always orthogonal to each other; (c) PCA transforms the data to smaller dimension that retains most of the information as a linear combination of the data, so we can find which features are important for best describing the variance in a broader sense; (d) PCA is an unsupervised learning method; (e) We did not assume so when deriving PCA.

3. In PCA, to find the principal components, we try to maximize _____:

- (a) the data likelihood
- (b) the variance of the data in the feature space
- (c) the variance of the data projected onto the principal components
- (d) the average norm of the data points projected onto the principal components
- (e) the approximation error of the data projected onto the principal components
- (f) the matrix norm of the data covariance matrix

Solution: (c). Please note that (e) is what PCA tries to minimize, and that (d) may also be true when data is zero-centered.

4. What kind of computation(s) are being done in PCA? Choose all that are correct.

- (a) $\max_{\mathbf{u}: \|\mathbf{u}\|_2=1} \frac{1}{N} \sum_n \mathbf{u}^\top \mathbf{x}^{(n)}$
- (b) $\max_{\mathbf{u}: \|\mathbf{u}\|_2=1} \frac{1}{N} \sum_n (\mathbf{u}^\top \mathbf{x}^{(n)} - \mathbf{u}^\top \bar{\mathbf{x}})^2$
- (c) $\max_{\mathbf{u}: \|\mathbf{u}\|_2=1} \frac{1}{N} \sum_n \|\mathbf{x}^{(n)} - (\mathbf{u}^\top \mathbf{x}^{(n)})\mathbf{u}\|^2$
- (d) $\min_{\mathbf{u}: \|\mathbf{u}\|_2=1} \frac{1}{N} \sum_n \|\mathbf{x}^{(n)} - (\mathbf{u}^\top \mathbf{x}^{(n)})\mathbf{u}\|^2$
- (e) Find the eigenvector(s) with the largest eigenvalue(s) of the data feature matrix
- (f) Find the eigenvector(s) with the largest eigenvalue(s) of the data covariance matrix
- (g) Find the eigenvector(s) with the smallest eigenvalue(s) of the data covariance matrix
- (h) Find the eigenvector(s) with the smallest eigenvalue(s) of the data covariance matrix

Solution: (b), (d), and (f).

(a) Note that $\bar{\mathbf{x}} \neq 0$ in general. (b) This is the variance maximization objective we've seen in the lecture. (d) This is the minimum distortion objective which is equivalent to the maximization objective; note that $(\mathbf{u}^\top \mathbf{x}^{(n)})\mathbf{u}$ is the projection of $\mathbf{x}^{(n)}$ onto \mathbf{u} .

5. In PCA, let's suppose we are finding the first principal component \mathbf{u}_1 and in order to do so you have found the (real) eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_R$ such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_R$, and their corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_R$ of the data covariance matrix. Then, how can you recover the projection variance for the first principal component?

Solution: λ_1 .

Note that the variance of the first principal component is same as $\mathbf{u}_1^\top \mathbf{S} \mathbf{u}_1$. This equals λ_1 because $\mathbf{S} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$.

EECS545 Lecture 17 Quiz Solutions

March 17, 2024

1. Select all that are true.

- (a) The goal of PCA is to interpret the underlying structure of the data in terms of the principal components that are best at predicting the output variable.
- (b) The number of source for the ICA can be larger than the input dimension.
- (c) The number of basis vectors for sparse coding can be larger than the input dimension.
- (d) tSNE is useful for pre-processing the data for downstream tasks (e.g., classification, regression, etc.)

Solution: (c)

2. Select all that are true about ICA.)

- (a) ICA can be used to reduce the dimensionality of data.
- (b) ICA can be used to discover independent components from data.
- (c) ICA finds directions (e.g., basis vectors) of maximal variance in the data.
- (d) ICA finds directions (e.g., basis vectors) that are maximally independent from data following a Gaussian distribution.
- (e) ICA finds directions (e.g., basis vectors) that are maximally independent from data following a non-Gaussian distribution.

Solution: (a) (b) (e)

3. What is the main benefit of applying PCA whitening before performing ICA on a dataset?

- (a) It reduces the dimensionality of the data by discarding low-variance components.
- (b) It makes the data more Gaussian by transforming it to a standard normal distribution.
- (c) It decorrelates the data by making its covariance matrix diagonal.
- (d) It rotates the data by aligning it with its principal directions.

Solution: (c)

PCA whitening decorrelates the data by making its covariance matrix diagonal. This simplifies the ICA problem by reducing it to finding a rotation matrix that maximizes non-Gaussianity.

4. Select all that are true about the sparse coding objective function.
- (a) Increasing β will result in a basis that better approximate the input data
 - (b) Using an L2 loss for the sparsity penalty is viable for sparse coding as long as β is increased
 - (c) The sparse coding objective is convex when maximizing b or s independently (while fixing the other), but not both.

Solution: (c)

Note for (b), L2 does not work well as a sparsity penalty (unlike L1, L0 and log penalty).

5. Select all that are true about tSNE vs. ISOMAP
- (a) ISOMAP preserves global distances while tSNE does not
 - (b) ISOMAP preserves local distance by preserving geodesic distance, while tSNE preserves (noisy) euclidean distance.
 - (c) tSNE and ISOMAP are both relatively computationally intensive.
 - (d) tSNE is less sensitive to noise than ISOMAP.

Solution: (b), (c), (d)

EECS545 Lecture 18 Quiz Solutions

1. The purpose of the generator G in GAN is to (check all that apply)
- (a) Maximize classification error for discriminator
 - (b) Minimize classification error for discriminator
 - (c) Minimize $\log(1 - D(G(z)))$.
 - (d) Maximize $\log(D(G(z)))$ as an approximate objective

Solution: (a), (c), (d) Note the GAN objective is:

$$\min_{\theta_g} \max_{\theta_d} [\mathbb{E}_{x \sim p_{\text{data}}} \log D_{\theta_d}(x) + \mathbb{E}_{z \sim p(z)} \log(1 - D_{\theta_d}(G_{\theta_g}(z)))]$$

where the term on the right is the term with the generated output for fake data.

2. Ideally, in GAN, when both generator and discriminator have enough capacity and training both of them converges after some iterations, which of the following statements are true? (Check all that apply.)
- (a) The discriminator can perfectly distinguish the samples from real data and the generated samples
 - (b) The generator can generate samples with the same distribution as the real data samples
 - (c) Both the generator and discriminator cannot be improved more
 - (d) The discriminator is unable to differentiate the real data samples and generated samples

Solution: (b), (c), (d)

3. Which of the following statements is true for a (vanilla) variational auto-encoder? (Check all that apply.)
- (a) VAE can be used for both continuous latent variables and discrete latent variables
 - (b) VAE efficiently approximates the maximum likelihood estimation for parameters in the network of encoder and decoder
 - (c) Training of VAE is more difficult than GAN
 - (d) VAE can be used to infer the latent variable z for input image x as latent representation, while GAN cannot be used to infer the latent representation.

Solution: (b), (d)

4. Will $\log(1 - D(G(z)))$ saturate early in learning if G is poor?

- (a) Yes, as discriminator can reject generated samples with high confidence
- (b) No, as discriminator can reject generated samples with high confidence
- (c) Yes, as discriminator will have very low confidence
- (d) No, as discriminator will have very low confidence

Solution: (a)

5. What are the differences between VAEs and diffusion models?

- (a) VAEs are inspired by non-equilibrium thermodynamics, while diffusion models rely on a surrogate loss
- (b) VAEs have access to a low-dimensional latent space, while diffusion models do not
- (c) VAEs can generate images in a single pass, while diffusion models have to generate through several passes
- (d) VAEs are approximate density models, while diffusion models are not

Solution: (b) and (c)

EECS545 Lecture 19 Quiz Solutions

1. Consider that you are trying to train an ML model, and the training loss does not seem to be dropping (assume we trained this model for a lot of epochs). Which of the following can be the possible reasons? (Choose all that apply)
- (a) Learning rate is too low.
 - (b) Weights are not properly initialized.
 - (c) The coefficient for the regularization term is too low.
 - (d) The dataset is too small.

Solution: (a), (b).
(c) and (d) would drop the learning rate faster than expected.

2. Which of the following methods can be used to reduce the high variance problem? (Choose all that apply)
- (a) Using dropout.
 - (b) Using the aggregated result from multiple models.
 - (c) Using a high coefficient for the regularization term while training.
 - (d) Changing the weight initialization.

Solution: (a), (b), (c).
Weight initialization (d) does not contribute to the high variance issues, in general.

3. Which of the following is true about ablative analysis? (Choose all that apply)
- (a) It generally tries to explain the difference between the poorer baseline and the current performance.
 - (b) It is performed by removing components from your system one at a time.
 - (c) It focuses more on explaining the difference between the current and perfect performance (100%) compared to the error analysis.
 - (d) It helps to understand the contribution of the component to the overall system.

Solution: (a), (b), (d).
(c): Error analysis focuses primarily on explaining the difference between the current performance and the perfect performance (100%).

4. (True/False) It is always possible to tell if an algorithm has converged and finished training by looking at the iteration/loss plot.

Solution: False. Your optimizer sometimes may train with a small learning rate in the beginning (e.g., warm-up when training a Transformer from scratch). Also, your objective sometimes does not well-matched with your problem. It is recommended to check whether you are optimizing the right function.

5. (True/False) There exists one optimization algorithm that would work on any problem. So, we do not need to change the optimizer from the beginning to the end for whatever problem settings.

Solution: False. You can start with a single optimizer, like ADAM, but it doesn't mean this optimizer will work for all problem settings. It is always recommended to check with multiple plausible options.

EECS545 Lecture 20 Quiz Solutions

1. Which of the following are true about Decision Trees? (Select all that apply)
- (a) Decision Trees are often simpler to understand and interpret than other models
 - (b) Decision Tree learning is stable under variations in data (e.g., sampling training data)
 - (c) Decision Trees can be used for multi-class classification problems
 - (d) Decision Trees are good for learning outputs that depend on multiple inputs, such as XOR, parity, or multiplexer problems

Solution: (a) and (c).

2. Which of the following are true about ensemble methods? (Select all that apply)
- (a) Using an Adaboost over a single weak learner reduces bias
 - (b) Using an averaging-based ensemble over a single model reduces variance
 - (c) Averaging-based ensemble methods are guaranteed to improve performance (e.g., generalization error) if individual models (weak classifiers) in the ensemble are independent.
 - (d) AdaBoost doesn't work well with depth-1 decision trees

Solution: (a) and (b).

3. Suppose we have two independent variables X and Y . What is the information gain $IG(X, Y)$?

Solution: 0. Intuitively, we cannot gain any information about X from Y (or vice-versa) because they are independent.

Notice that $H(X) = H(X|Y = y) = H(X|Y)$ because $p(x|y) = p(x)$. Hence, $IG(X, Y) = H(X) - H(X|Y) = 0$.

4. Suppose we roll two fair 6-sided dice, dice A and dice B. Suppose the sum of the two dice is 4. What is the information gain on dice A if we know the value of dice B? I.e. What is $IG(A, B)$?

Solution:

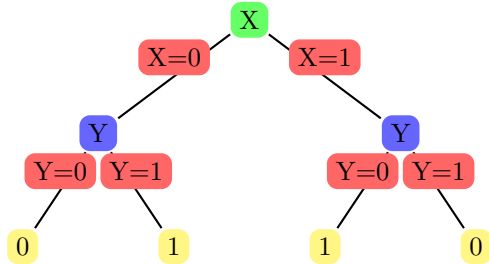
$$H(A) = -(1/3) * \log(1/3) - (1/3) * \log(1/3) - (1/3) * \log(1/3)$$

$H(A|B)$ must be 0, because the value of B completely determines the value of A.

$$IG(A|B) = H(A) = 1.585$$

5. Suppose we would like to solve the XOR problem with a decision tree. I.e. Given $X, Y \in \{0, 1\}$, we would like to predict $\text{XOR}(X, Y)$. What is the minimum number of nodes in the decision tree needed to predict XOR with 100% accuracy? Include leaf nodes in your count.

Solution: 7. We will need the maximum possible of nodes for a 2-feature decision tree. 1 root node, 2 middle-nodes, and 4 leaf nodes.



This tikz picture was draw using ChatGPT (with some light editing)!

Queries: “Can you draw me a tikz picture of a decision tree classifying the XOR problem?” Output:

```
\begin{tikzpicture}[grow=right,level distance=1.5cm]
\tikzstyle{every node}=[fill=red!60,rounded corners]
\tikzstyle{edge from parent}=[red,-o,thick,draw]
\tikzstyle{edge from parent path}=[red,-o,thick,draw]
\tikzstyle{level 1}=[sibling distance=4cm,thick]
\tikzstyle{level 2}=[sibling distance=2cm,thick]
\node [fill=green!60]{XOR}
child {node [fill=blue!60] {AND}
child {node [fill=yellow!60] {0}}
child {node [fill=yellow!60] {1}}
}
child {node [fill=blue!60] {AND}
child {node [fill=yellow!60] {1}}
child {node [fill=yellow!60] {0}}
};
\end{tikzpicture}
```

I further queried ChatGPT to add edges, and relabel some nodes. Amazing!

EECS545 Lecture 20 Quiz Solutions

1. Given a sequence of states [Sunny, Sunny, Rainy, Sunny, Rainy, Rainy, Sunny, Sunny, Sunny] and a sequence of observations (precipitation in inches) [0, 0, 0.2, 0, 0.1, 0.3, 0, 0, 0], estimate the observation probability distribution of precipitation on Rainy days $P(\text{precipitation}|\text{Rainy}) = \mathcal{N}(\mu, \sigma^2)$.

Solution:

$$\begin{aligned}\mu &= \text{Sample mean of precipitation when it is Rainy} \\ &= \frac{1}{3}(0.2 + 0.1 + 0.3) \\ &= 0.2 \\ \sigma &= \text{Sample standard deviation of precipitation when it is Rainy} \\ &= \sqrt{\frac{1}{3-1}((0.2 - \mu)^2 + (0.1 - \mu)^2 + (0.3 - \mu)^2)} \\ &= \sqrt{\frac{1}{2}((-0.1)^2 + 0.1^2)} \\ &= 0.1\end{aligned}$$

2. Continued. Estimate the probability of transitioning Sunny \rightarrow Rainy.

Solution: We use Bayes Rule.

$$P(\text{Sunny} \rightarrow \text{Rainy}) = \frac{\# \text{ Sunny} \rightarrow \text{Rainy}}{\# \text{ Sunny} \rightarrow \text{Sunny} + \# \text{ Sunny} \rightarrow \text{Rainy}} = \frac{2}{5}$$

3. Which of the following statements are true about HMMs and RNNs? (Select all that apply)
- (a) HMM models a hidden state in data but RNN does not
 - (b) HMM models an emission probability from hidden state but RNN does not
 - (c) For HMM, you are guaranteed to converge to a local optimum, and it requires less tuning than RNNs.
 - (d) HMM can model the data with richer set of hidden states and dynamics.
 - (e) HMMs require fewer data to train

Solution: (c) and (e).

4. True or False. When learning an HMM for a fixed set of observations, assume we do not know the true number of hidden states (which is often the case), we can always monotonically increase the training data likelihood by permitting more hidden states.

Solution: True. Intuitively, if we increase the number of hidden states to the extreme, a hidden state for every timestep, then we can “fit” that the transition and emission probabilities perfectly for every state, making the likelihood 1! Increasing the number of hidden states in general similarly increases the likelihood but at a smaller scale.