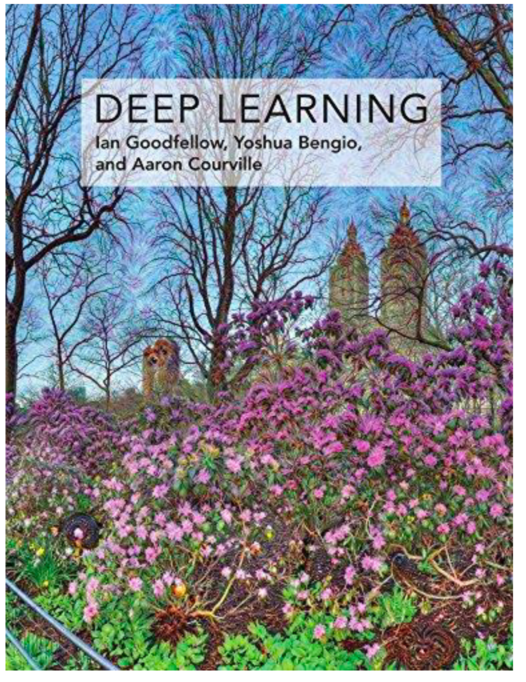
CMPS 392

Final Exam
Thursday, May 14, 2020 4:00 PM
Prepared by: Mohamed Nassar



Q1

- The following equation: $p(y = 1|\boldsymbol{x}); \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$ corresponds to:

  a) Linear regression

  b) Logistic regression

  c) Decision tree

  d) Support Vector Machines

  e) A neural network with one hidden layer

(Goodfellow 2016)

Q2

- Finding a linear function f such that $f([0,1],w) = 1$ and $f([1,0],w) = 1$ but $f([1,1],w) = 0$ and $f([0,0],w) = 0$ is called:

a) The OR problem

b) The NAND problem

c) The XOR problem

d) The perceptron

(Goodfellow 2016)

Q3

- **The drosophila** of machine learning is:
  a) The Fashion MNIST dataset
  b) The MNIST dataset
  c) The Grocery dataset
  d) The CIFAR-10 dataset
  e) The CIFAR-100 dataset

(Goodfellow 2016)

Q4

- **Underfitting** occurs when:
  1. The model is not able to obtain a sufficiently low error value on the training set.
  2. The gap between the training error and test error is too large.
  3. The model has low capacity
  4. The model has high capacity
  5. The model is not a neural network

(Goodfellow 2016)

Q5

- **Gradient descent** has the advantage to:
  1. Get attracted to saddle points
  2. Avoid computing the Hessian
  3. Not to get attracted to saddle points
  4. Get stuck in local minima
  5. Get stuck in large flat regions

(Goodfellow 2016)

Q6

- When the condition number of the Hessian is large,
  a) It is better to keep the learning rate small
  b) It is better to keep the learning rate large
  c) It does not matter
  d) It is better to compute the optimal step size
  e) It makes the gradient descent optimization harder
  f) It makes the gradient descent optimization easier

(Goodfellow 2016)

Q7

- What is numerically stable solution to $\log\left(\frac{e^{x_i}}{\sum e^{x_j}}\right)$:

a) $\log\left(\frac{e^{x_i-m}}{\sum e^{x_j-m}}\right)$

b) $\log\left(\frac{e^{x_i}}{\sum e^{x_j-m}}\right)$

c) $\log\left(\frac{e^{x_i-m}}{\sum e^{x_j}}\right)$

d) $\frac{1}{m}\log\left(\frac{me^{x_i}}{\sum e^{x_j}}\right)$

(Goodfellow 2016)

Q8

- The **no Free lunch** theorem means that:
  a) parametric models are better than non-parametric modelds.
  b) the manifold hypothesis is not always true.
  c) machine learning problems become exceedingly difficult when the number of dimensions in the data is high.
  d) no machine learning algorithm is universally any better than any other machine learning algorithm.

(Goodfellow 2016)

Q9

- The kernel trick in SVM is:
  a) To replace $\boldsymbol{x}$ by $\phi(\boldsymbol{x})$
  b) To replace $\phi(\boldsymbol{x_i})\phi(\boldsymbol{x_j})$ by $k(\boldsymbol{x_i}, \boldsymbol{x_j})$
  c) To maximize the width of the street
  d) To define the constraints as $y_i(\boldsymbol{w}\boldsymbol{x_i} + b) \geq 1$

(Goodfellow 2016)

Q10

- PCA finds a transformation $\boldsymbol{z} = \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{x}$ where:
  a) The covariance of $\boldsymbol{z}$ is diagonal
  b) $\boldsymbol{W}$ is an orthogonal matrix
  c) A column in $\boldsymbol{W}$ is an eigen vector of $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$
  d) Taking $\boldsymbol{W}_{\boldsymbol{l}}$ composed of the first l columns of $\boldsymbol{W}$ minimizes the loss $\left\|\boldsymbol{x} - \boldsymbol{W}_{\boldsymbol{l}}\boldsymbol{W}_{\boldsymbol{l}}^{\boldsymbol{T}} x\right\|_2$

(Goodfellow 2016)

Q11

- *k*-means clustering:
  a) Starts with one cluster, then checks if it can be divided into two, and repeat.
  b) Starts with a predefined number of clusters
  c) Starts with considering each point as a cluster, then checks if two clusters can be merger together
  d) Is a supervised learning algorithm

(Goodfellow 2016)

Q12

- The following equation represents:

$$J(\boldsymbol{w}, b) = -\mathbb{E}_{\mathbf{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}} \log p_{\mathrm{model}}(y \mid \boldsymbol{x})$$

a) Cross-entropy

b) Mean squared error

c) Reconstruction error

d) Negative log likelihood

(Goodfellow 2016)

Q13

- The smoothness prior:
  a) Is also known as local consistency
  b) It means that $f(x+\epsilon) \approx f(x)$
  c) It assumes that most of $\mathbb{R}^n$ consists of invalid inputs
  d) It means that interesting inputs occur only along a collection of manifolds

(Goodfellow 2016)

Q14

- If we assume that the probability distribution of the output of a neural network follows a Bernoulli distribution, which output layer and cost function are more convenient?
  a) Sigmoid and binary cross-entropy
  b) Softmax and discrete cross-entropy
  c) Linear and mean squared error
  d) Sigmoid and mean squared error

(Goodfellow 2016)

Q15

- If $q_i = \frac{\exp(z_i)}{\sum \exp(z_j)}$ and $J = -\sum p_j \log(q_j)$, what is $\frac{\partial J}{\partial z_i}$?

a) $\frac{\partial J}{\partial z_i} = \sum_j \frac{\partial J}{\partial q_j} \frac{\partial q_j}{\partial z_i}$

b) $\frac{\partial J}{\partial z_i} = \left(\sum_{j \neq i} \frac{\partial J}{\partial q_j} \frac{\partial q_j}{\partial z_i}\right) + \frac{\partial J}{\partial q_i} \frac{\partial q_i}{\partial z_i}$

c) $\frac{\partial J}{\partial z_i} = \left(\sum_{j \neq i} -\frac{p_j}{q_j} \frac{\partial q_j}{\partial z_i}\right) - \frac{p_i}{q_i} \frac{\partial q_i}{\partial z_i}$

d) $\frac{\partial J}{\partial z_i} = \left(\sum_{j \neq i} -\frac{p_j}{q_j} (-q_j q_i)\right) - \frac{p_i}{q_i} \left(q_i (1 - q_i)\right)$

e) $\frac{\partial J}{\partial z_i} = \left(\sum_{j \neq i} p_j q_i\right) - p_i (1 - q_i)$

(Goodfellow 2016)

Q16

- If $q_i = \frac{\exp(z_i)}{\sum \exp(z_j)}$ and $J = -\sum p_j \log(q_j)$, what is $\frac{\partial J}{\partial z_i}$?

a) $q_i - p_i$

b) $p_i - q_i$

c) $p_i \log(q_i)$

d) $q_i \log(p_i)$

(Goodfellow 2016)

Q17

- In L^2 Parameter regularization:
  a) Weights are projected into unit L2 ball
  b) Weights are shrunk by a multiplicative factor
  c) Weights are shifted by an additive factor
  d) Some weights are nullified, the others are kept the same

(Goodfellow 2016)

Q18

- L1 regularization:
  a) Encourages sparsity
  b) Encourages small weights
  c) equivalent to MAP Bayesian estimation with Gaussian prior
  d) equivalent to MAP Bayesian estimation with Laplace prior

(Goodfellow 2016)

Q19

- Which statements are true about dropout?
  a) Dropout is equivalent to bagging of several independent models.
  b) Dropout is equivalent to bagging of several models sharing parameters.
  c) Each model is trained to convergence on its respective training set.
  d) A tiny fraction of the models are each trained for a single step.
  e) The training set encountered by each sub-network is a subset of the original training set sampled with replacement.

(Goodfellow 2016)

Q20

- Which of the following are regularization techniques?
  a) Early-stopping
  b) Nesterov Momentum
  c) Adversarial training
  d) Information erasing
  e) AdaGrad
  f) Dataset augmentation
  g) RMSProp
  h) Training with added noise

(Goodfellow 2016)

Q21

- Which statements are true about batch normalization?
  a) It is a good way to isolate the updates across many layers
  b) It requires to compute the mean of each unit for a batch of activations
  c) It is only used at training time
  d) It requires to compute the standard deviation of each unit for a batch of activations
  e) It does not require any learnable parameters

(Goodfellow 2016)

Q22

- Some of the convolutional neural networks' properties are:
  a) dense connections
  b) sparse connections
  c) Weight sharing
  d) Weight scaling
  e) Equivariance to translation
  f) Equivariance to rotation

(Goodfellow 2016)

Q23

- The pooling operation helps to:
  a) Make the representation invariant to small translations of the input
  b) Make the representation more efficient
  c) Make the deconvolution operation straightforward
  d) Avoid the need for backpropagation

(Goodfellow 2016)

Q24

- Valid padding model is when:
  a) The convolution kernel is only allowed to visit positions where the kernel is contained entirely within the image
  b) We pad with enough zeroes to preserve the input dimension
  c) We pad with enough zeroes to make every input contribute to equal number of outputs

(Goodfellow 2016)

Q25

- Audio data that has been transformed with a Fourier transform to a matrix of amplitudes where rows correpsond to frequencies and columns correspond to different points in time:
  a) Have two dimensions and one channel
  b) Have one dimension and two channels
  c) using CNN, we preserve equivariance to a shift in octaves
  d) Using CNN, we preserve equivariance to a shift in amplitude

(Goodfellow 2016)

Q26

- A recurrent neural network:
  a) Maps an arbitrary length sequence $\boldsymbol{x}^{\boldsymbol{t}}, \boldsymbol{x}^{\boldsymbol{t-1}}, \boldsymbol{x}^{\boldsymbol{t-2}}, \ldots, \boldsymbol{x}^{\boldsymbol{2}}, \boldsymbol{x}^{\boldsymbol{1}}$ to a fixed length vector $\boldsymbol{h}^{t}$
  b) Is not different than 1D CNN
  c) Produces each member of the output using the same update rule applied to the previous outputs
  d) It is based on sharing local parameters within a very small neighborhood using a kernel function

(Goodfellow 2016)

Q27

- An LSTM cell:
  a) Has an internal recurrence and an external recurrence
  b) Has an output unit that can be shut off by the output gate
  c) Has three gates in addition to the input and the output units
  d) Has an input gate which is equivalent to the identity function
  e) Has a state unit that can be nullified by a forget gate
  f) In a LSTM layer, is connected to all the other cells

(Goodfellow 2016)

Q28

- Logistic regression:
  a) Is equivalent to $\sigma(x^T w)$
  b) The cost function is $-\log\left(\sigma(x^T w)\right)$ if $y = 1$
  c) The cost function is $-\log\left(1 - \sigma(x^T w)\right)$ if $y = 0$
  d) The cost function is $-\log\left(\sigma(-x^T w)\right)$ if $y = 0$
  e) The cost function is $-\log\left(\sigma((2y - 1)x^T w)\right)$
  f) The cost function is $\zeta\left((1 - 2y)x^T w\right)$
  g) The gradient is $\sigma\left((1 - 2y)x^T w\right)(1 - 2y)x$
  h) The gradient is $(\sigma(x^T w) - 1)x$ if $y = 1$
  i) The gradient is $\sigma(x^T w)x$ if $y = 0$
  j) The gradient is $(\sigma(x^T w) - y)x$

(Goodfellow 2016)