CS/DS 541: Class 23

Jacob Whitehill

Final Exam

Final Exam Topics

- Fundamentals
  - Jacobian matrices
  - Hessian matrices
  - Multivariate chain rule of differential calculus
  - Probability theory
- Loss functions
  - MSE
  - Log loss
  - Cross-entropy
- Shallow models
  - Linear regression
  - Logistic regression
  - Softmax regression

Final Exam Topics

- Deep models
  - Feed-forward neural networks
  - Universal function approximation theorem
  - Recurrent neural networks
  - Convolutional neural networks
  - Auto-Encoders
  - Generative-Adversarial Networks
  - Variational Auto-Encoders

Final Exam Topics

- Optimization methods and issues
  - Back-propagation
  - (Stochastic) gradient descent
  - Newton's method
  - Momentum
  - ADAM, Adagrad, RMSProp
  - Ill-conditioned loss functions
  - Vanishing & exploding gradients
- Regularization methods
  - L1 regularization
  - L2 regularization
  - Dropout
  - Data augmentation
  - Ensembles
- Feature normalization methods
  - Whitening
  - Batch normalization

Exercises

https://cs230.stanford.edu/files/cs230exam_win19.pdf
https://cs230.stanford.edu/files/cs230exam_win20.pdf

Exercise 1

(2 points) You have an input volume of $32 \times 32 \times 3$. What are the dimensions of the resulting volume after convolving a 5×5 kernel with zero padding, stride of 1, and 2 filters? (1 formula)

Exercise 2

You're asked to build an algorithm estimating the risk of premature birth for pregnant women using ultrasound images.

(2 point) You have 500 examples in total, of which only 175 were examples of preterm births (positive examples, label = 1). To compensate for this class imbalance, you decide to duplicate all of the positive examples, and then split the data into train, validation and test sets. Explain what is a problem with this approach. (1-2 sentences)

Exercise 3

(1 point) You fix the issue. Subject matter experts tell you that the model should absolutely not miss preterm births, but false positives are okay. Your best model achieves 100% recall. Does it mean the model works well? Explain. (1 sentence)

Exercise 4

(1 point) You are training a standard GAN, and at the end of the first epoch you take note of the values of the generator and discriminator losses. At the end of epoch 100, the values of the loss functions are approximately the same as they were at the end of the first epoch. Why are the quality of generated images at epoch 1 and epoch 100 not necessarily similar? (1-2 sentences)

Exercise 5

(1 point) What would you set the padding of a 2D CONV layer to be (as a function of the filter width f) to ensure that the output has the same dimension as the input? Assume the stride is 1. (1 formula)

Exercise 6

(1 point) Recall that a neural network with a single hidden layer is sufficient to approximate any continuous function (with some assumptions on the activation). Why would you use neural networks with multiple layers? (1 sentence)

Exercise 7

Many supermarket customers use the yellow creamy spot on the outside of a watermelon to evaluate its level of sweetness. To help customers who aren't aware of this fact, you decide to build an image classifier to predict whether a watermelon is sweet (label=1) or not (label=0).

(a) **(1 point)** You've built your own labeled dataset, chosen a neural network architecture, and are thinking about using the mean squared error (MSE) loss to optimize model parameters. Give one reason why MSE might not be a good choice for your loss function. (1 sentence)

Exercise 8

(1 point) You decide to use the binary cross-entropy (BCE) loss to optimize your network. Write down the formula for this loss (for a single example) in terms of the label y and prediction $\hat{y}$. (1 formula)

Exercise 9

You decide to train one model with L2 regularization (model A) and one without (model B). How would you expect model A's weights to compare to model B's weights? (1 sentence)

Exercise 10

The price of a watermelon depends on its weight, rather than its level of sweetness. Thus supermarkets don't care about a watermelon's level of sweetness as much as customers do.

Supermarkets give you a new dataset of watermelon images and their corresponding weight in pounds, and ask you to build another image classifier to predict the weight of a watermelon.

(2 point) You decide to use a single unified neural network to predict both the level of sweetness and the weight of a watermelon given an image.

Propose a new loss function to train the unified model. Assume no regularization and write your answer in terms of the new y and $\hat{y}$. (1 formula)

Exercise 11

Two historians approach you for your deep learning expertise. They want to classify images of historical objects into 3 classes depending on the time they were created:

- Antiquity ($y = 0$)
- Middle Ages ($y = 1$)
- Modern Era ($y = 2$)

(A) Class: Antiquity

(B) Class: Middle Ages

(C) Class: Modern Era

Figure 1: Example of images found in the dataset along with their classes

Over the last few years, the historians have collected nearly 5,000 hand-labelled RGB images.

(i) **(2 points)** Before training your model, you want to decide the image resolution to be used. Why is the choice of image resolution important?

Exercise 12

(1 point) If you had 1 hour to choose the resolution to be used, what would you do?

Exercise 13

(3 points) As you train your model, you realize that you do not have enough data. Cite 3 data augmentation techniques that can be used to overcome the shortage of data.

Exercise 14

(1 point) Consider a Generative Adversarial Network (GAN) which successfully produces images of apples. Which of the following propositions is **false**?

(i) The generator aims to learn the distribution of apple images.

(ii) The discriminator can be used to classify images as apple vs. non-apple.

(iii) After training the GAN, the discriminator loss eventually reaches a constant value.

(iv) The generator can produce unseen images of apples.

Exercise 15

(2 points) Consider a simple convolutional neural network with one convolutional layer. Which of the following statements is true about this network? (Check all that apply.)

(i) It is scale invariant.

(ii) It is rotation invariant.

(iii) It is translation invariant.

(iv) All of the above.

Exercise 16

(2 points) Mini-batch gradient descent is a better optimizer than full-batch gradient descent to avoid getting stuck in saddle points.

(i) True

(ii) False

Exercise 17

Consider an input image of shape $500 \times 500 \times 3$. You flatten this image and use a fully connected layer with 100 hidden units.

(i) **(1 point)** What is the shape of the weight matrix of this layer?