Quizzes

Week 6

Q1. Artificial and biological neurons behave identically.

False! False! A million times false! Artificial neurons (and neural networks) are inspired by biological neurons but have very very different behavior and computation.

Q2.A single neuron with a logistic activation can represent non-linear boundaries in classification problems. False. A single neuron computes a linear function of its inputs and passes it through a logistic function-matching logistic regression.

Q3. What is a loss function?

A loss function is a function that measures how bad a network's output is, usually relative to some gold-standard for what the output should be.

Q4. Networks with Sigmoid activations tend to converge faster than ReLU networks.

False. ReLUs tend to converge faster because they do not saturate for positive inputs like Sigmoid.

Week 7

Q1. Explain the benefits of backpropagation.

Backpropagation is an efficient way to compute gradients of the loss with respect to model parameters. Firstly, it is a reverse-mode differentiation and computes the product of intermediate Jacobians from the output of the network backwards -- reducing cost of matrix multiplication in computing gradients. Secondly and more important, the backwards pass allows us to store and reuse loss gradients as we work our way backwards through the network.

Q2. What is a Jacobian?

We refer to the gradient of vector-valued functions with respect to vector inputs as a Jacobian matrix. If y(x) is a function that outputs a vector given an input vector x, then the Jacobian dy(x)/dx would be a matrix where the i,j'th entry would be the partial derivative of the i'th element of y(x) with respect to the j'th element of x.

Q3. Neural networks are universal approximators regardless of their size.

False. For any given function and error tolerance, there exists a neural network of finite size that achieves that level of error; however, a fixed-size network may or may not be large enough for any given function.

Q4. What is a computational graph and how can it be used to implement backpropagation?

A computational graph is a directed acyclic graph with node corresponding to units of computation and edges corresponding to the results of these computations. If each node implements both a forward and backward (i.e. computing a Jacobian of output with respect to input) operation, then the graph can be used to calculate derivatives of the output of arbitrary combinations of operations via backpropagation.

Q5. Convolutional neural networks make the assumption that relevant features in an input can be found by examining local regions of the input -- e.g. windows of time in a sequence or regions of an image.

True! The locality assumption is a key part of convolutional neural networks.

Q6.Decision trees are composed of internal nodes that provide label outputs and leaf nodes that implement tests to divide datapoints.

False. The other way around -- internal nodes implement tests to divide datapoints and leaf nodes are labels.

Q7. Splits in standard decision trees can be diagonal lines.

False. In standard decision trees, the splits are axis-aligned because they test only single attributes.

Q8.Finding a decision tree with minimum depth that achieves minimum error on training data has polynomial computational complexity. False. This problem is NP-Hard because all possible combinations of tree structures need to be considered.

Week 8

Q1. Decision trees have a tendency to overfit training data. What are options for reducing overfitting?

- Limiting the depth of the tree
- Monitor the validation accuracy and perform early stopping
- Removing nodes based on validation performance from the tree after training has finished

Q2. Explain how decision trees can deal with continuous valued attributes.

Even though continuous valued attributes have an infinite number of possible thresholds, only a finite set of thresholds need to be considered during training because thresholds occurring between the same datapoints will have identical purity measures like Information Gain. As such, we can just consider thresholds at the midpoint between sequential datapoints for an attribute.

Q3. Decision tree learning can terminate whenever all attributes result in zero information gain.

False. We discussed this as a potential base case; however, many functions (like XOR) may not show any information gain in initial variables but be able to be usefully split later only after multiple variables have been considered.

Q4.Bias -> Error due to assumptions in the model not matching the problem (aka modelling error)**Variance** -> Error due to sensitivity to changes in the dataset (aka estimation and optimization error)**Weak Learner** -> Models that tend to have high bias but low variance.**Strong Learner** -> Models that tend to have low bias but high variance

Q5. Given a training dataset {a, b, c, d, e}, show three example bootstrap training sets that could be used for training bagged models.

Each training dataset for bagging is a sample with replacement from the original dataset with the same number of instances. So I drew a random number between 1 and 5 five times to make each dataset below: {d, d, b, a, d} {b, d, a, d, b} {a, c, e, b, a}

Q6. Boosting is a method to make an ensemble of weak learner stronger -- i.e., a way to reduce bias in weak learners. Boosting typically trains a sequence of models with each model focusing on the errors of the previous models.

True! We talked about two versions of this -- L2 boosting and Adaboost

Q7.Bagging is an ensemble method for reducing the variance of weak learners by combining multiple models trained on random samples of the dataset. False. Bagging is used to reduce variance in **strong learners**. Weak learners already have low variance, but high bias.