CS 446: Machine Learning
Discussion Session

Daniel Khashabi

October 20, 2015

1 Neural Networks

1.1 true of false

- Backpropagation algorithm always achieve the optimal solution
- Number of nodes in hidden layer can control generalization
- Weights in neural networks have intuitive meaning
- Neural networks can be used for interpolating a function.
- Convergance is guaranteed in Backpropagation algorithm.

1.2 Backpropagation

Consider a neural net for a binary classification which has one hidden layer as shown in the Figure Figure 1. We use a linear activation function $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ at the hidden units, and a sigmoid activation function $g(\mathbf{z}) = \frac{1}{1+e^{\mathbf{w}^\top \mathbf{z}}}$ at the output unit.

- Find the value of the hidden variables z_1 and z_2, given x_1 and x_2.
- Find the value of the output variable y, given hidden variables y_1 and y_2.
- For fixed weights, under what input values, the network will predict +1 in the output?
- Define an error function E to be squared loss (the error between the predictions and target values). Find the gradient of the error with respect to the weights incoming to the output layer, i.e. $\frac{\partial E}{\partial w_i}$, for $w_i \in \{w_7, w_8, w_9\}$.
- Find the gradient of the error with respect to the weights incoming to the hidden layer, i.e. $\frac{\partial E}{\partial w_i}$, for $w_i \in \{w_1, w_2, w_3, w_4, w_5, w_6\}$.
- Given a training instance $((\hat{x}_1, \hat{x}_2), \hat{y})$, what are the Backpropagation update rules for one iteration using this training instance.

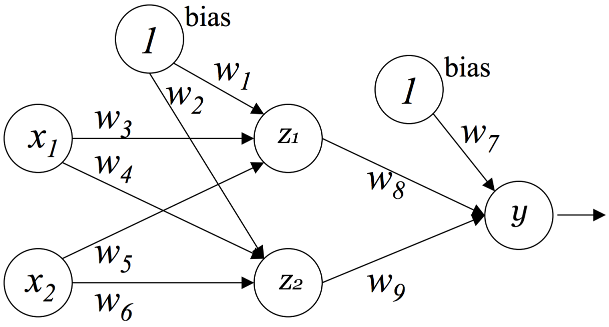

Figure 1: A two-layer neural network.

2 decision tree (true or false)

- Suppose the value of an attribute is the same across all of the training instances. Would removing this attribute change the resulting decision tree?
- Suppose we have repetition in the training set (the input-output pairs are exactly the same). Would removing the repetitions change the resulting decision tree?
- The ID3 algorithm is guaranteed to find the optimal decision tree.

3 Fitting data into a model (beyond linear regression)

1. Consider the following function:

$$f(x) = ax^2 + bx, \quad x \in \mathbb{R}$$

   Given training data consisting of the input-output pairs $D = \{(x_1, y_1), ..., (x_n, y_n)\}$, find the parameters a and b such that best fit the training data, by minimizing the sum-of-squared loss functions.

2. Given limited data, would a linear model better fit the "train" data, or the quadratic model in the previous part? How about the "test" data?

4 Learning theory

1. Which of the following procedures is sufficient and necessary and most efficient for proving that the VC dimension of a learner is N?

   (a) Show that the classifier can shatter all possible dichotomies with N points.

   (b) Show that the classifier can shatter all possible dichotomies with N points.

   (c) Show that the classifier can shatter a subset of all possible dichotomies with N points.

   (d) Show that the classifier can shatter all possible dichotomies with N points and that it cannot shatter any of the dichotomies with N+1 points.

(e) Show that the classifier can shatter all possible dichotomies with N points and that it cannot shatter one of the dichotomies with N+1 points.

(f) Show that the classifier can shatter a subset of all possible dichotomies with N points and that it cannot shatter one of the dichotomies with N+1 points.

2. Find the VC-dimension of the following: $f(w^\top w + \theta)$, where f is an arbitrary increasing non-linear function ($w, x \in \mathbb{R}^d$ and $\theta \in \mathbb{R}$)

3. For the concept class in the previous question, find the minimum number of training instances (sample complexity) necessary to learn a hypothesis with error at most ϵ with probability at least $1 - \delta$.

4. **True of False?:** AdaBoost will eventually reach zero training error, regardless of the type of weak classifier it uses, provided enough weak classifiers have been combined.

   Answer: False! If the data is not separable by a linear combination of the weak classifiers, AdaBoost cannot achieve zero training error.