CS 475 Machine Learning: Homework 6
Structured Prediction
Due: Wednesday December 7, 2016, 11:59pm

Jonathan Liu
jliu118

1 Analytical (25 points)

1. (12 points) One of the uses of PCA is to create better data representations for classification. Consider two different classes in a binary classification task. The first class has points generated using a Gaussian with $\mu_1 = \{1, -1\}$ and covariance

$$\mathbf{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & .001 \end{pmatrix}$$

The second class has points generated using a Gaussian with $\mu_2 = \{1, 1\}$ and covariance

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & .002 \end{pmatrix}$$

a. How would a linear classifier perform when trained to classify between these two classes in the given representation? Why?

b. Suppose PCA is run on this two dimensional data to produce a one dimensional representation. Describe the principal component that PCA would select.

c. How would a linear classifier perform when trained to classify between these two classes in the one dimensional PCA representation? Why?

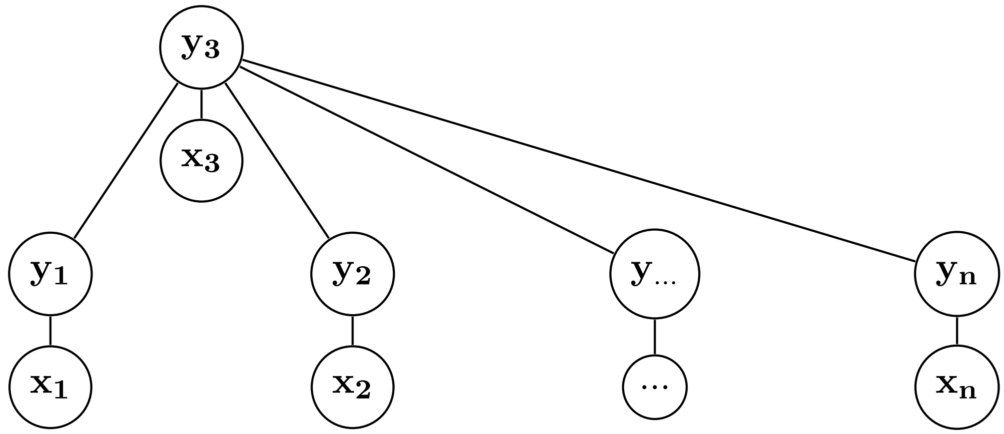
d. Suppose instead you train a neural network on this dataset. The neural network has a single hidden layer with a single hidden node, and a single output node corresponding to the label. Compare the performance of this neural network to a linear classifier trained on a representation of this data learned by PCA.

a. A linear classifier would perform perfectly because there is very little variance in the x_2 dimension, and a lot of variance in the x_1 dimension, so since there's only a lot of variance in one dimension, a linear classifier would do really well.

b. It would select x_1 as the principal component because it has the highest variance in both of the classes.

c. It would be difficult to linearly separate the two classes in the one dimensional PCA representation because since we are choosing x_1 as our principal component, the mean for both classes is one. And if they are centered around the same location, they would be hard to separate using a linear classifier.

d. A neural network would perform better than a linear classifier on a representation of this data learned by a PCA. This is because using a neural network, we can ignore the x_1 dimension and work only with x_2 dimensions which is easier to separate because the means are more different and lower variance.



2. (13 points) Consider the graphical model shown above. In this model, $\mathbf{x}$ is a sequence of observations for which we want to output a prediction $\mathbf{y}$, which itself is a sequence, where the size of $\mathbf{y}$ is the same as $\mathbf{x}$. Unlike sequence models we discussed in class, this model has a tree structure over the hidden nodes. Assume that the potential functions have a log-linear form: $\psi(Z) = \exp\{\sum_i \theta_i f_i(Z)\}$, where Z is the set of nodes that are arguments to the potential function (i.e. some combination of nodes in $\mathbf{x}$ and $\mathbf{y}$,) θ are the parameters of the potential functions and f_i is a feature function.

a. Write the log likelihood for this model of a single instance $\mathbf{x}$: $p(\mathbf{y}, \mathbf{x})$.

b. Write the conditional log likelihood for this model of a single instance $\mathbf{x}$: $p(\mathbf{y}|\mathbf{x})$.

c. Assume that each variable y_i can take one of k possible states, and variable x_i can take one of k' possible states, where k' is very large. Describe the computational challenges of modeling $p(\mathbf{y}, \mathbf{x})$ vs $p(\mathbf{y}|\mathbf{x})$.

d. Propose an efficient algorithm for making a prediction for $\mathbf{y}$ given $\mathbf{x}$ and θ.

a.

$$\log(p(y,x)) = (\sum_i \log(\psi((x_i, y_i))) + \sum_i \log(\psi((y_i, y_3)))) - \log(Z)$$

$$\log(Z) = \sum(\sum_i \log(\psi((x_i, y_i))) + \sum_i \log(\psi((y_i, y_3))))$$

because this is log, we convert to multiplication and subtraction.

b.

c.

d.