Name: ______________________________ NetID: ____________

STATISTICS AND DATA SCIENCE 355 / 555

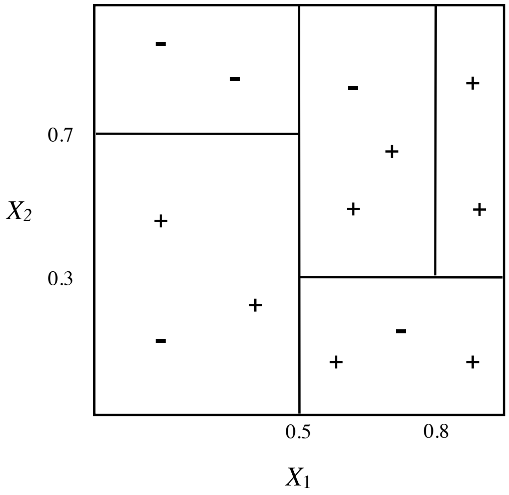
Introductory Machine Learning

Quiz 2 (practice), Thursday, October 31, 2019

1. *Decision trees* (5 points)

   Consider the following figure showing 13 points in $\mathbb{R}^2$ and a partition of the unit square $[0, 1] \times [0, 1]$. Eight points are from class $Y = 1$ (labeled "+") and five points are from class $Y = -1$ (labeled "-").

draw tree below:



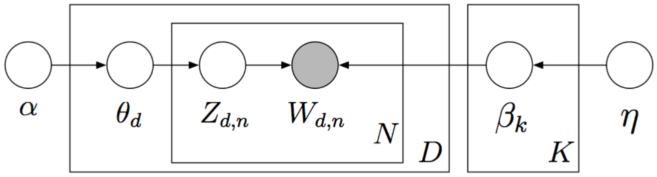
(a) Draw (to the right of the above figure) the decision tree that corresponds to the illustrated partition. Label the questions asked at each node.

(b) What is the training error (as a percentage) for these 13 data points?

(c) What is the predicted value of Y for the point $X = (X_1, X_2) = (.6, .2)$?

2. *Topic modeling* (5 points)

The latent Dirichlet allocation topic model is represented by the diagram



where $\theta_d \sim$ Dirichlet(α) are the per-document topic proportions, $Z_{d,n} \sim$ Multinomial(θ_d) are the per-word topic assignments, $W_{d,n} \sim$ Multinomial$(\beta_{Z_{d,n}})$ are the observed words, and $\beta_k \sim$ Dirichlet(η) are the topics.

Circle the correct answers:

TRUE FALSE (1) The model is generative, and can assign a probability to documents that are not in the training data.

TRUE FALSE (2) According to the model, each document is generated by a single topic.

TRUE FALSE (3) According to the model, the words are generated independently.

TRUE FALSE (4) As α decreases from one toward zero, the topic proportions vector θ_d tends to have small values for a larger number of topics.

TRUE FALSE (5) The Gibbs sampling algorithm chooses the most probable topic $Z_{d,n}$ for a selected word $W_{d,n}$ while holding all of the other Z values fixed.