CS 6375
Prof. Nick Ruozzi
Student: Tri Minh Cao
Date: September 30, 2017

Problem Set 2

Problem 1: Breast Cancer Diagnosis

Part 1. Primal SVMs

Accuracy on training set and validation set:

C	Training set accuracy	Validation set accuracy
1	0.5289	0.5271
10	0.5289	0.5247
10e2	0.5323	0.5247
10e3	0.5303	0.5247
10e4	0.531	0.5176
10e5	0.531	0.5176
10e6	0.531	0.5176
10e7	0.531	0.5176
10e8	0.531	0.5176

Best value for c: c = 1.

Using c = 1. Accuracy on test set: 0.5593

Part 2. Dual SVMs with Gaussian Kernels

C	Sigma	Training set accuracy	Validation set accuracy
1	0.1	1.0	0.6
1	1	0.997	0.6
1	10	0.8929	0.8941
1	100	0.881	0.9059
1	1000	0.8512	0.8824
10	0.1	1.0	0.6
10	1	1.0	0.5882
10	10	0.9613	0.8706
10	100	0.881	0.9059
10	1000	0.869	0.8824
10e2	0.1	1.0	0.6
10e2	1	1.0	0.5882
10e2	10	0.9821	0.8824
10e2	100	0.875	0.8941
10e2	1000	0.8452	0.8588
10e3	0.1	1.0	0.6
10e3	1	1.0	0.5882
10e3	10	0.994	0.8706
10e3	100	0.9077	0.8941
10e3	1000	0.8512	0.8588
10e4	0.1	1.0	0.6
10e4	1	1.0	0.5882
10e4	10	1.0	0.8471
10e4	100	0.9137	0.9294
10e4	1000	0.744	0.8
10e5	0.1	1.0	0.6
10e5	1	1.0	0.5882
10e5	10	1.0	0.8471
10e5	100	0.8363	0.8471

10e5	1000	0.8482	0.8353
10e6	0.1	1.0	0.6
10e6	1	1.0	0.5882
10e6	10	1.0	0.8471
10e6	100	0.8571	0.8235
10e6	1000	0.8869	0.8824
10e7	0.1	1.0	0.6
10e7	1	1.0	0.5882
10e7	10	1.0	0.8471
10e7	100	0.8244	0.7765
10e7	1000	No solution*	No solution
10e8	0.1	1.0	0.6
10e8	1	1.0	0.5882
10e8	10	1.0	0.8471
10e8	100	No solution	No solution
10e8	1000	No solution	No solution

*In some cases of c and sigma, the quadratic solver (cvxpy + ECOS) could not find a solution.

Best value for c and sigma: c = 10000; sigma = 100

Using c = 10000 and sigma = 100. Accuracy on test set: 0.8784.

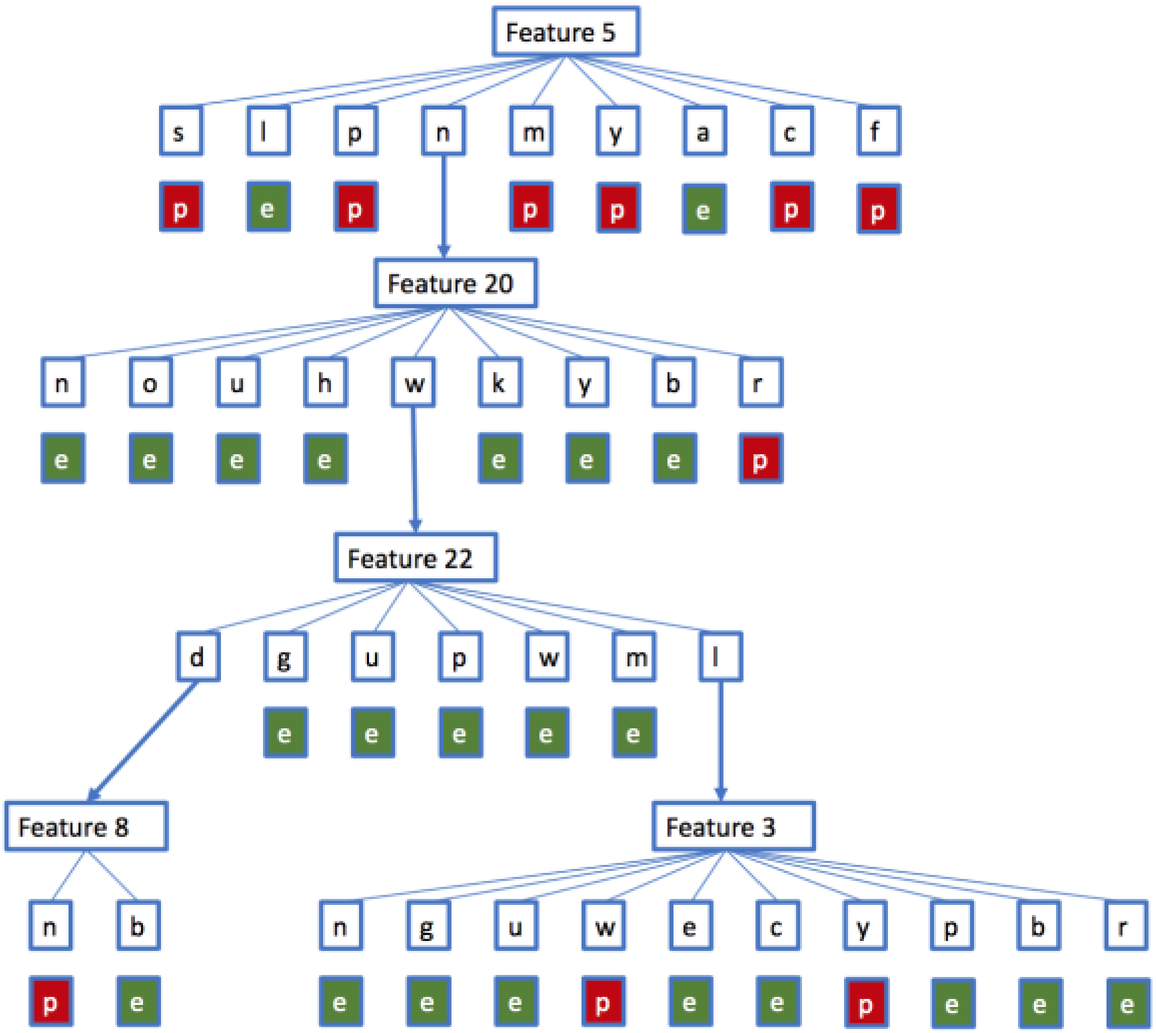
Part 3. K-Nearest Neighbor

K	Test set accuracy
1	0.9527
5	0.9595
11	0.973
15	0.973
21	0.98

4. For this classification task, k-nearest neighbor is considerably better than SVM. Not only kNN gives better classification rate (0.98 vs 0.878 on test set), it is also much faster (for this problem).

Problem 2: Poisonous Mushrooms?

1. The learned decision tree:



2. Size of the decision tree: 37 nodes

3. Height of the decision tree: 4

4. Accuracy on training set: 1.0

5. Accuracy on test set: 1.0

6. Decision tree works very well for this problem. At least with the data we have, the decision tree is very confident about classifying edible from poisonous mushrooms. I think one reason that the Society Field Guide does not want to give a simple set of rules is the risk inherent in eating wild mushrooms. If the mushroom is indeed poisonous, then the risk is too high to try eating.

7. The quality of the learned decision tree is dependent on the training/test split.
Decision tree is inherently overfitting so if the training set does not have data about some feature, then the learned decision tree will not be able to make a decision on that feature.
Suppose a feature *i* has 4 possible values: a, b, c, d. But in the training set, we only observe values a and b from feature *i*. In that case, if the test set has samples with values c and d for feature *i*, the learned decision tree cannot decide what to do with those values.

8. Yes. For this problem, the best decision tree with exactly one non-leaf node is equal to the one found by using information gain to select one attribute. That attribute is feature 5.