**Problem 1: Spam, Spam, Ham**

1. Primal SVMs

|  |  |
| --- | --- |
| c | Accuracy on training set |
| 1 | 93.4420 |
| 10 | 93.6957 |
| 100 | 93.7681 |
| 1000 | 93.6594 |
| 10000 | 93.6594 |
|  |  |

|  |  |
| --- | --- |
| c | Accuracy on validation set |
| 1 | 93.8043 |
| 10 | 94.1304 |
| 100 | ***94.4565*** |
| 1000 | 94.3478 |
| 10000 | 94.378 |
|  |  |

As we can see the maximum accuracy on the validation set is when c is 100. We select this c and check the accuracy of the test set.

|  |  |
| --- | --- |
| C | Accuracy on test set |
| 100 | 93.8111 |

1. DUAL SVMs with GAUSSIAN KERNELS

**Accuracy of the Training Set** on the learned classifier on different pairs of C and σ.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| C ↓ σ→ | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
| 1 | 100 | 100 | 99.9638 | 99.6377 | 84.6739 | 76.6304 |
| 10 | 100 | 100 | 100 | 99.9275 | 67.5725 | 73.7681 |
| 100 | 100 | 100 | 100 | 99.9638 | 62.7899 | 63.8768 |
| 1000 | 100 | 100 | 100 | 99.9638 | 62.8986 | 63.4420 |
| 10000 | 100 | 100 | 100 | 99.9638 | 55.7971 | 69.8188 |

**Accuracy of the Validation Set** on the learned classifier on different pairs of C and σ

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| C ↓ σ→ | 0.001 | 0.01 | 0.1 | 1 | 10 | 100 |
| 1 | 64.2391 | 66.0870 | 70.2174 | 77.6087 | ***84.4564*** | 76.0870 |
| 10 | 64.2391 | 66.0870 | 70.6522 | 77.9348 | 67.8261 | 73.2609 |
| 100 | 64.2391 | 66.0870 | 70.6522 | 77.8261 | 63.6957 | 63.5870 |
| 1000 | 64.2391 | 66.0870 | 70.6522 | 77.8261 | 63.3696 | 63.0435 |
| 10000 | 64.2391 | 66.0870 | 70.6522 | 77.6087 | 54.7826 | 69.0217 |

As we can see the maximum accuracy on the validation set is when c is 1 and σ is 10. We select this values of c and σ to train our classifier and check the accuracy on test set.

|  |  |
| --- | --- |
| C ↓ σ→ | 10 |
| 1 | 82.9533 |

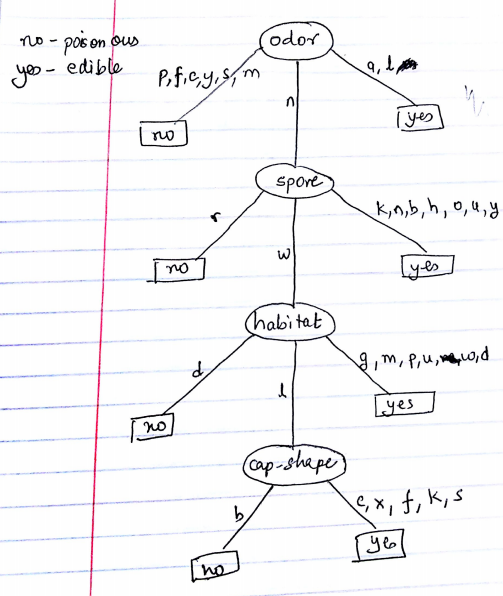
Please find the attached Matlab files for both of the above problems.

1. Which approach should be preferred?

If we answer this question based on the results above, Primal SVM with slack would be a better approach to spam-email classification task.

For spam email classification there can be some attributes that have more weightage than other attributes, so classifying this way where we consider all the attributes same is not a better classifier.

**Problem 2: Poisonous Mushrooms?**



The learned Decision Tree

1. The size of the learned decision tree is the count of the node here (not including leaf as they are class labels) is **4**.
2. The depth of the learned decision tree is 4.
3. Accuracy of the learned decision tree on training data set is 100%.
4. Accuracy of the learned decision tree on test data is 99.35285%.
5. The Audubon Society Field Guide to North American Mushrooms states that there is not a simple set of rules to determine whether or not a mushroom is edible. The decision tree approach for this problem to determine whether or not a mushroom is edible is not perfect. Let’s say we learn a decision tree taking every features of mushrooms present today, now if tomorrow a new species of mushroom is found that is not classified by the decision tree, using majority vote would not give an accurate classification.
6. We split the data set into train and test so as to calculate the accuracy of the learned decision tree. If the training set is huge (nearly 100%) there will be very less data in the test set to check its accuracy. Whereas if the training set is very less (nearly 0%) there will be too much of data in the test set and very few to construct the classifier. Therefore there should be an appropriate split on the training/test data such that there are enough samples in train to get the best classifier and enough samples in test to check the accuracy.

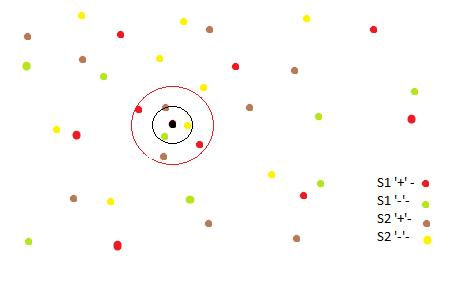
**Problem 3: Understanding nearest Neighbor Methods**

1. Let there be two distinct training sets S1 and S2 where labels are + or -.

Now we have to classify a new point x using 1-nearest neighbor algorithm. If 1-nearest neighbor algorithm gives + using S1 only and it gives + using S2 only.

If we take the union of S1 and S2 and use 1-nearest neighbor algorithm the classification will be + as the nearest point from the new data point x will be either from S1 or S2. Thus the 1-nearest neighbor algorithm will label x with a + when using S1 union S2 as the training set.

1. Now if we use the 3-nearest neighbor algorithm the classification will not be + necessary as there can be two minuses out of the three nearest neighbors. See the example below



Union of S1 and S2, two concentric circles to see the nearest points. We can see that 3-nearest neighbor for union gives – but for individual S1 and S2 we are getting +.