# Support Vector Machine (SVM); Classification Task; Example in Python; AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Please include (1) Python file; (2) Output from Python with answers to all the questions asked; (3) Comparison of the results to those using R; (4) Recommended websites for SVM using Python.

#### SVM with the Titanic Data – Classification Task

The Titanic.csv data we will use for our homework is taken from the Kaggle competition site (<https://www.kaggle.com/c/titanic>) where it was called the train.csv. **We will treat this dataset as our entire data** because we do not know the survival status in the Kaggle test.csv data. Our Titanic data has 891 passengers and 12 variables:

* *PassengerId*: Passenger ID: 1- 891
* *Survived*: A binary variable indicating whether the passenger survived or not (0 = No; 1 = Yes); this is our response variable
* *Pclass*: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd)
* *Name*: A field rich in information as it contains title and family names
* *Sex*: male/female
* *Age*: Age, a significant portion of values are missing
* *SibSp*: Number of siblings/spouses aboard
* *Parch*: Number of parents/children aboard
* *Ticket*: Ticket number.
* *Fare*: Passenger fare (British Pound).
* *Cabin*: Cabin number
* *Embarked*: Port of embarkation (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)

First, one must clean the data and decide which variables to exclude from our analysis. My recommendation is that we exclude *PassengerId, Name*, *Ticket*, and *Cabin* in the ensuing analysis. Next, please note that *Age* has many missing values – my suggestion is to delete those with missing values. Now after the data cleaning step, your task is to split the data randomly into training (75%) and testing (25%), first build the best SVM models to predict passenger survival using the training data, and then use these models to predict whether each passenger in the testing data survived or not. Please note that we use the ***caret***package in R to build the various SVM classifiers.

**Note:** For this data set, we shall only perform data standardization for the variables *Age* and *Fare*.

Please review the following website for related methods and procedures in R:

<http://www.sthda.com/english/articles/36-classification-methods-essentials/144-svm-model-support-vector-machine-essentials/>

1. For the entire dataset, please perform the data cleaning as instructed before; namely, exclude the variables *Name*, *Ticket*, and *Cabin* and delete missing values in the variable *Age*. Please normalize the data as instructed. (For this data set, we shall only perform data standardization for the variables *Age* and *Fare*.) Please also let R know that both Survived and PClass are categorical variables although they appear to have numerical values. Please report how many passengers are left after this step. Then please use the random seed 123 to divide the cleaned data into 75% training and 25% testing.

1. Please first build the best classifier to predict passenger survival using the training data and the linear SVM. We shall use the default value for the cost parameter C. Please compute the Confusion matrix and report the sensitivity (that is, a passenger who survived is predicted to have survived), specificity (that is, a passenger who did not survive is predicted to not have survived), and the overall accuracy using the testing data. Please add the predicted class label to the testing dataset.
2. Next we will build the best classifier to predict passenger survival using the training data and the linear SVM. We shall find the optimal cost parameter C by using the command line:

tuneGrid = expand.grid(C = seq(0, 2, length = 20))

Please (i) report the optimal cost parameter value, and (ii) compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data.

1. Now we shall build the best classifier to predict passenger survival using the training data and the SVM with radial basis kernel. We shall find the optimal tuning parameters C and sigma (**σ**) by using the command line:

tuneLength = 10

Please (i) report the optimal parameter values, and (ii) compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data.

1. Now we shall build the best classifier to predict passenger survival using the training data and the SVM with polynomial basis kernel. We shall find the optimal tuning parameters C, degree and scale by using the command line:

tuneLength = 4

Please (i) report the optimal parameter values, and (ii) compute the Confusion matrix and report the sensitivity, specificity, and the overall accuracy using the testing data.

1. Which SVM classifier, namely, (1) linear SVM, (2) SVM with radial basis kernel, and (3) SVM with polynomial kernel, give us the best overall accuracy for the Titanic data?

