**Code book**

**1. Study design**

This project uses data from the Human Activity Recognition Using Smartphones Dataset. The data come from an experiment, which has been performed with 30 subjects, performing a number of activities wearing a smartphone (Samsung Galaxy S II) on the waist. The activities performed were: WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers were selected for generating the training data and 30% the test data.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain.

**2. Variables**

The obtained features he features selected for this database come from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Similarly, the acceleration signal was then separated into body and gravity acceleration signals (tBodyAcc-XYZ and tGravityAcc-XYZ) using another low pass Butterworth filter with a corner frequency of 0.3 Hz.

Subsequently, the body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyAccJerk-XYZ and tBodyGyroJerk-XYZ). Also the magnitude of these three-dimensional signals were calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag).

Finally a Fast Fourier Transform (FFT) was applied to some of these signals producing fBodyAcc-XYZ, fBodyAccJerk-XYZ, fBodyGyro-XYZ, fBodyAccJerkMag, fBodyGyroMag, fBodyGyroJerkMag. (Note the 'f' to indicate frequency domain signals).

These signals were used to estimate variables of the feature vector for each pattern:

'-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.

The set of variables that were estimated from these signals are:

mean(): Mean value

std(): Standard deviation

mad(): Median absolute deviation

max(): Largest value in array

min(): Smallest value in array

sma(): Signal magnitude area

energy(): Energy measure. Sum of the squares divided by the number of values.

iqr(): Interquartile range

entropy(): Signal entropy

arCoeff(): Autorregresion coefficients with Burg order equal to 4

correlation(): correlation coefficient between two signals

maxInds(): index of the frequency component with largest magnitude

meanFreq(): Weighted average of the frequency components to obtain a mean frequency

skewness(): skewness of the frequency domain signal

kurtosis(): kurtosis of the frequency domain signal

bandsEnergy(): Energy of a frequency interval within the 64 bins of the FFT of each window.

angle(): Angle between to vectors.

Additional vectors obtained by averaging the signals in a signal window sample. These are used on the angle() variable:

gravityMean

tBodyAccMean

tBodyAccJerkMean

tBodyGyroMean

tBodyGyroJerkMean

**3. Data transformation incorporated in this project**

Refer to the script (run\_analysis.R) for more details on each task.

The test folder (containing x\_test, y\_test, and subject\_test) were combined to obtain the complete test data. Similarly, training data was combined and the two sets were pooled together (merged).

**3.1 Extracting mean and standard deviations**

From the combined data sets, variables were selected that contained the mean and standard deviation of a measurement, plus the activity type (ID), and subject ID.

**3.2 Use descriptive activity names**

To identify the activities, a new column was added to the data set, which has details on the activity type (called activity ID in this case). This was done using the grepl logical operator and subsetting from the original data only those columns that contained the words “mean” and “std”.

**3.3 Label variables appropriately**

Some shortcuts and abbreviations from the names of the variables were removed, using information in the features\_into.txt file (function gsub was used for that).

**3.4 Tidy data set**

The latter data set was then transformed to present the means of all variables by activity type and subject. The (plyr) packages was used with the function ddply which created averages by subjectId and activityID for all columns, i.e. all variables (using coluwise()).