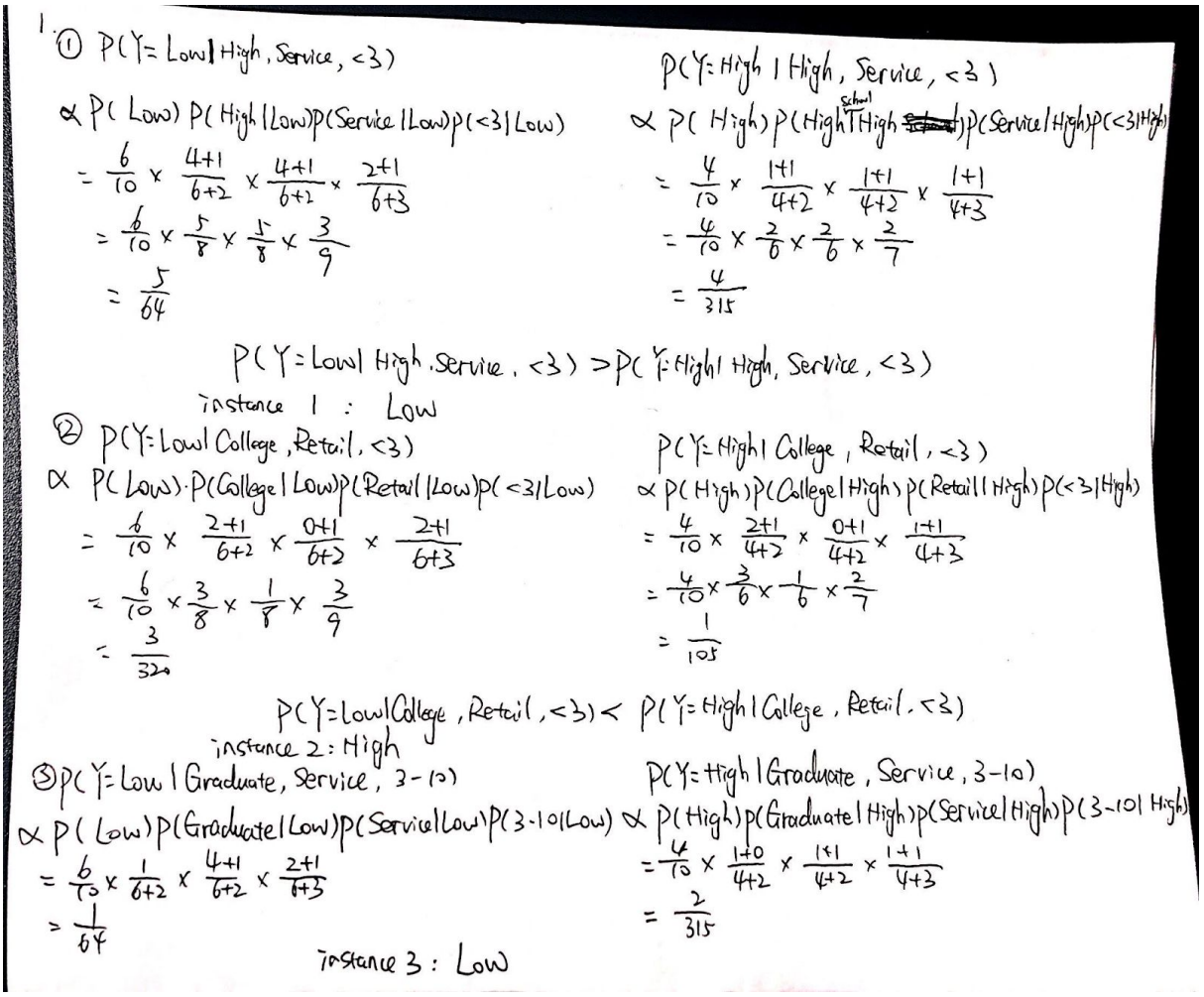
1.



2.

(a) List the features from highest |r| (the absolute value of r) to lowest, along with their |r| values. Why would one be interested in the absolute value of r rather than the raw value?

[('f4', 0.43692179751745097),
('f13', 0.36826904080902556),
('f14', 0.36822372149726035),
('f16', 0.36602511423650724),
('f7', 0.35214126136392476),
('f22', 0.3513499255347562),
('f26', 0.341042614938154),
('f1', 0.308810814577297),
('f20', 0.29904900743176754),
('f31', 0.29078291134679674),
('f34', 0.2660927897329379),
('f2', 0.1957323905545247),
('f28', 0.15690433267657294),
('f25', 0.1530959899437491),
('f19', 0.13763622058924627),
('f17', 0.11394472976390747),
('f32', 0.09317373256743067),
('f8', 0.08777300962449965),

('f0', 0.06979505192021568),
('f10', 0.056876488922882024),
('f21', 0.05660516824019749),
('f11', 0.04211688398659001),
('f33', 0.03880964782127297),
('f6', 0.03529477533787),
('f15', 0.03147794480573731),
('f35', 0.030855237164366982),
('f29', 0.020829454125099515),
('f18', 0.01793142539798069),
('f27', 0.015606234757601551),
('f9', 0.013005370998714192),
('f3', 0.009213581453788188),
('f30', 0.008955194801128104),
('f24', 0.0077797430847418745),
('f23', 0.005507866072800373),
('f12', 0.0021785840278091897),
('f5', 9.813778651783369e-05)]

If the correlation between feature and label is negative, it will be sorted as the smaller one than 0 if we just use r. But actually, the negative correlation also reflects their strong relation between feature and label. So we need to include negative correlations by absolute r value.

(b) Select the features that have the highest m values of |r|, and run LOOCV on the dataset restricted to only those m features. Which value of m gives the highest LOOCV classification accuracy, and what is the value of this optimal accuracy?
features:
('f4', 'f13', 'f14', 'f16', 'f7', 'f22', 'f26', 'f1', 'f20', 'f31', 'f34', 'f2', 'f28', 'f25', 'f19', 'f17', 'f32', 'f8', 'f0', 'f10')
accuracy:
0.925531914893617

3. (40 points) Wrapper Method Starting with the empty set of features, use a greedy approach to add the single feature that improves performance by the largest amount when added to the feature set. This is Sequential Forward Selection. Define performance as the LOOCV classification accuracy of the KNN classifier using only the features in the selection set (including the ?candidate? feature). Stop adding features only when there is no candidate that when added to the selection set increases the LOOCV accuracy.

(a) Show the set of selected features at each step, as it grows from size zero to its final size (increasing in size by exactly one feature at each step).
['f20']
['f20', 'f10']
['f20', 'f10', 'f19']
['f20', 'f10', 'f19', 'f8']
['f20', 'f10', 'f19', 'f8', 'f7']

['f20', 'f10', 'f19', 'f8', 'f7', 'f14']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f1']
['f20', 'f10', 'f19', 'f8', 'f7', 'f14', 'f2', 'f1', 'f16']

(b) What is the LOOCV accuracy over the final set of selected features?

0.9657210401891253