# MET CS 677 A3

# Assignment 5

**Question 1:**

1. **download IRIS dataset, remove Setosa flowers and assign labels 0 to to Versicolor and 1 to Virginica.**
2. **for each label and feature compute statistical averages (from training set!) and put them in the following table:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | μ0 | σ0 | μ1 | σ1 | μall | σall |
| **Sepal Length** | 5.94 | 0.52 | 6.59 | 0.64 | 6.26 | 0.66 |
| **Sepal Width** | 2.77 | 0.32 | 2.97 | 0.32 | 2.87 | 0.33 |
| **Petal Length** | 4.26 | 0.47 | 5.55 | 0.55 | 4.91 | 0.83 |
| **Petal Width** | 1.33 | 0.2 | 2.03 | 0.27 | 1.68 | 0.42 |

Table 1: Statistical measurements

YYYY

1. **for each class, compute the correlation matrix for your 4 features. Which features have the highest and lowest cor­relations?**

For Setosa:

The highest cor­relations of sepal-width and sepal-length is 0.746780.

The lowest cor­relations of petal-length and sepal-width is 0.176695

-----------------------------------------------------------------------------------------

For Versicolor:

**Answer:**

**highest cor­relations**: petal-width and petal-length is 0.786668

The lowest cor­relations of sepal-length and sepal-width is 0.525911

-----------------------------------------------------------------------------

For Virginica

Answer:

The highest cor­relations of petal-length and sepal-length is 0.864225

The lowest cor­relations of petal-width and sepal-length is 0.281108

1. **discuss your findings**

Answer:

For all the classes,

Virginica’s petal-length and sepal-length has the highest correlation ,0.864225.

Setosa’s petal-length and sepal-width has the lowest correlation, 0.176695

**Question 2:**

**1.generate histograms of pairwise relationships for a train­ing set (include these histograms in submitted homework). X\_train. You can use "pairplot” method of the seaborn package:**

**2.examine the histograms and for each feature design a simple classifier ("weak learner") for labels. Your classifier can only consist of simple comparison using that single feature.**

Classifier for sepal-length: If > 5, label = 0, else = 1

Classifier for sepal-width: If > 3, label = 0, else = 1

Classifier for petal-length: If > 4, label = 0, else = 1

Classifier for petal-width: If > 2, label = 0, else = 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | **TP** | **TN** | **FP** | **FN** | **Accuracy** |
| (1) sepal length | 0 | 24 | 3 | 23 | 48% |
| (2) sepal-width | 13 | 4 | 23 | 10 | 34% |
| (3) petal-length | 0 | 19 | 8 | 23 | 38% |
| (4) petal-width | 12 | 0 | 27 | 11 | 24% |

**3. discuss your findings and rank your "weak" learners by ac­curacy (from most accurate to least accurate)**

Answer:

Rank: sepal-length > petal-length > sepal-width > petal-width

Sepal length has the most accurate Confusion Matrix

**Question 3:**

**1.For each such ensemble classifier, split data into training and test. Apply your classifiers on testing data, compute confusion matrix and summarize the results in a table below (note that no training is done, we are just combining the "weak” learners).**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | **TP** | **TN** | **FP** | **FN** | **Accuracy** |
| (1) (2)(3) | 0 | 24 | 3 | 23 | 48% |
| (1) (2)(4) | 23 | 0 | 27 | 0 | 46% |
| (1) (3)(4) | 2 | 1 | 26 | 21 | 6% |
| (2) (3)(4) | 23 | 0 | 27 | 0 | 46% |

Table 3: Result for ensembles of "weak'' learners

**2.discuss your findings and rank your ensembles learners by accuracy (from most accurate to least accurate**

**Answer:**

**Rank:** Sepal-length > sepal-width = petal-width > petal-length

Sepal-length has the most accurate Confusion Matrix

**3.compare "weak learners”" and ensemble results.**

**Answer:** Compare to weak learners, the ensemble results are the same with it, the Sepal-length is more accurate

**Question 4:**

**1.you design 4 such density-based classfiers, one for each of the 4 features. For each classifier, compute the confusion matrix (from a testing set! as before) and summarize them in a table below**

Classifier for sepal-length: If (p\_0 >= p\_1), label = 0,

else = 1

Classifier for sepal-width: If (p\_0 <= p\_1), label = 0,

else = 1

Classifier for petal-length: If (p\_0 - p\_1) >= 0, label = 0,

else = 1

Classifier for petal-width: If (p\_0 - p\_1) <= 0, label = 0,

else = 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | **TP** | **TN** | **FP** | **FN** | **Accuracy** |
| (1) sepal length | 18 | 21 | 6 | 5 | 78% |
| (2) sepal-width | 8 | 26 | 1 | 15 | 68% |
| (3) petal-length | 4 | 27 | 0 | 19 | 62% |
| (4) petal-width | 0 | 27 | 0 | 23 | 54% |

Table 4: Results for Density-based "weak" learners

**2.discuss your findings and rank your density-based "weak" learners by accuracy (from most accurate to least accurate**

**Answer:**

**Rank:** Sepal-length > sepal-width > petal-length > petal-width

Sepal-length has the most accurate Confusion Matrix

**Question 5:**

**1.For each such ensemble classifier, compute confusion matrix (on testing data!) and summarize the results in a table below**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Classifier | **TP** | **TN** | **FP** | **FN** | **Accuracy** |
| (1) (2)(3) | 23 | 0 | 27 | 0 | 46% |
| (1) (2)(4) | 23 | 0 | 27 | 0 | 46% |
| (1) (3)(4) | 23 | 0 | 27 | 0 | 46% |
| (2) (3)(4) | 23 | 0 | 27 | 0 | 46% |

Table 5: Results for Density-based Ensembles

**2.discuss your findings and rank your ensembles learners by accuracy (from most accurate to least accurate)**

**Answer:**

**Rank:** Sepal-length = sepal-width = petal-length = petal-width

They have the same accurate Confusion Matrix

**3.compare "weak learners, and ensemble results.**

**Answer:**

Compared to weak learners, the conditions which used to label them are conflict with each other, so the matrix may be not shown correctly. According to the classifier, the value of the accuracy are different.

**Question 6:**

**1.give a quick summary on comparing classifiers in Method I and Method II**

**Answer:**

Compared to Method 2, I think Method 1 is more accurate than Method 2, because we can not ignore that the conflict with each conditions, if the conditions are conflict with each other, it will cause the huge change for the result.