Introduction to machine learning Exercise set 4 report

Elias Annila 014328901

Pen and paper:

1.

Intuitively it would seem that set A is more pure as it has larger fraction of Audis, the dominant class, on the other hand set B has clearer difference between Volvos and Toyotas. As in the slides let k be number of classes and p_i fraction representing class i

Entropy:

$$E(A)=-\sum_{i=0}^{k}(p_i\log_2 p_i)=-(\frac{10}{80}\log_2\frac{10}{80}+\frac{25}{80}\log_2\frac{25}{80}+\frac{45}{80}\log_2\frac{45}{80})=1.37$$

$$E(B)=-\sum_{i=0}^{k}(p_i\log_2 p_i)=-(\frac{8}{80}\log_2\frac{8}{80}+\frac{32}{80}\log_2\frac{32}{80}+\frac{40}{80}\log_2\frac{40}{80})=1.36$$

Gini:

$$G(A)=1-\sum_{i=0}^{k}p_i^2=1-(\frac{10^2}{80^2}+\frac{25^2}{80^2}+\frac{45^2}{80^2})=0.57$$

$$G(B)=1-\sum_{i=0}^{k}p_i^2=1-(\frac{8^2}{80^2}+\frac{32^2}{80^2}+\frac{40^2}{80^2})=0.58$$

Classification error:

$$C(A)=1-max\,p_i=1-\frac{45}{80}=0.44$$

$$C(B)=1-max\,p_i=1-\frac{40}{80}=0.5$$

According to these impurity measures it is not clear which set is more pure. Classification error was in agreement with intuition of larger majority as expected, and Gini impurity also would suggest set A to be more pure. Entropy however suggests set B to be more pure.

For a binary classification such disagreement is not possible. For a binary case the measures are:

$$E_b(p)=-(p\log_2 p+(1-p)\log_2(1-p))$$
$$G_b(p)=1-(p^2+(1-p)^2)$$
$$C_b(p)=1-p\;when\,p>0.5$$
$$C_b(p)=p\;when\,p<0.5$$

p is the fraction of + cases.

If a disagreement between the three measures were to occur it would require there to be fractions p_i and p_j for which one of the measures would produce higher impurity for p_i and another for p_j.

Considering p_i is smaller than p_j it would mean that when going from p_i to p_j one of the functions had to rise and another to descend. So we look at each impurity measure's derivate and their zero points:

Entropy:

$$E_b(p)=-(p\log_2 p+(1-p)\log_2(1-p))$$
$$E_b(p)=-\frac{p\ln p}{\ln 2}-\frac{(1-p)\ln(1-p)}{\ln 2}$$
$$E_b{}'(p)=-\frac{\ln p+p\frac{1}{p}}{\ln 2}-\frac{D[1-p]\ln(1-p)+(1-p)D[\ln(1-p)]}{\ln 2}$$
$$E_b{}'(p)=-\frac{\ln p+1}{\ln 2}-\frac{-\ln(1-p)+(1-p)\frac{1}{1-p}(-1)}{\ln 2}$$
$$E_b{}'(p)=-\frac{\ln p+1}{\ln 2}-\frac{-\ln(1-p)-1}{\ln 2}$$
$$E_b{}'(p)=\frac{-\ln p+\ln(1-p)}{\ln 2}$$

$$E_b{}'(p)=0$$
$$-\ln p+\ln(1-p)=0$$
$$\ln p=\ln(1-p)|0\le p\le 1$$
$$p=1-p$$
$$p=0.5$$

$$E_b{}'(0.1)=\frac{-\ln 0.1+\ln(0.9)}{\ln 2}=1.5$$
$$E_b{}'(0.6)=\frac{-\ln 0.6+\ln(0.4)}{\ln 2}=-0.9$$

from this we know that between 0 and 0.5 E_b is rising and between 0.5 and 1 it's descending.

Gini:

$$G_b(p)=1-(p^2+(1-p)^2)$$
$$G_b(p)=1-p^2-(1-p)^2$$
$$G_b(p)=1-p^2-1+2p-p^2$$
$$G_b(p)=-2p^2+2p$$

$$G_b{}'(p)=-4p+2$$
$$-4p+2=0$$
$$4p=2$$
$$p=0.5$$

as G_b clearly is downward opening parabel and $G_b' = 0$ at 0.5 we can see that G_b is rising between 0 and 0.5 and descending between 0.5 and 1

Classification error:

$$C_b(p)=-p \, when \, p>0.5$$
$$C_b(p)=p \, when \, p<0.5$$

from this it is obvious that C_b is rising between 0 and 0.5 and descending between 0.5 and 1.

As all three functions are always rising and descending at same values of p it is not possible that there would be a disagreement between any of them regarding which is more pure p_i or p_j.

2.

I was a bit unsure if an instance had to have all of the attributes or if it could only have any combination of attributes, here I am presuming that each instance must have all attributes.

a)

The rule set is not consistent as for example in a case where an instance has following attributes: Airconditioner = Broken and Mileage = Low, regardless of the other attributes the first rule outputs high value and last rule low value.

b)

The rule set is complete as last three rules always lead to a decision based on Air conditioning and Engine, regardless of other attributes.

c)

The meaning is not clear as there are rules that may classify an instance to both low or high value classes for example instance with high mileage and working air conditioning and good engine.

d)

Yes the output of the rule list depends on the order of the rules. For example for the instance presented above changing first rule to last would result in classification changing from low to high.

e)

No there is no need for a default rule as all possible instances get classified to either high or low. The rule set is already complete without adding default rule.

3.

Minimizing classification error is equal to minimizing loss in 0-1 loss situation so optimal classifier

$$f_*(x)=arg\,max\,(P(y|x))$$

$$f_*(x)=arg\,max(\frac{P(y)P(x|y)}{P(x)})$$

$P(x)$ is constant and $P(y=0)=P(y=1)$

$$f_*(x)=arg\,max\,(P(x|y))$$

so as optimal classifier depends only on class conditional distributions we can choose decision boundary based on normal curve probability densities:

$$P(x|y=0)>P(x|y=1)$$

$$\frac{e^{\frac{-x_1^2}{2\sigma_0^2}}}{\sigma_0^2\sqrt{2\pi}}\frac{e^{\frac{-x_2^2}{2\sigma_0^2}}}{\sigma_0^2\sqrt{2\pi}}>\frac{e^{\frac{-x_1^2}{2\sigma_1^2}}}{\sigma_1^2\sqrt{2\pi}}\frac{e^{\frac{-x_2^2}{2\sigma_1^2}}}{\sigma_1^2\sqrt{2\pi}}$$

$$\frac{e^{\frac{-x_1^2}{2\sigma_0^2}}e^{\frac{-x_2^2}{2\sigma_0^2}}}{(\sigma_0^2)^2}>\frac{e^{\frac{-x_1^2}{2\sigma_1^2}}e^{\frac{-x_2^2}{2\sigma_1^2}}}{(\sigma_1^2)^2}$$

$$e^{\frac{-x_1^2}{2}}e^{\frac{-x_2^2}{2}}>\frac{e^{\frac{-x_1^2}{32}}e^{\frac{-x_2^2}{32}}}{256}$$

$$\ln(256)-\frac{x_1^2}{2}-\frac{x_2^2}{2}>\frac{-x_1^2}{32}-\frac{x_2^2}{32}$$

$$32\ln(256)-16x_1^2-16x_2^2>-x_1^2-x_2^2$$

$$32\ln(256)>15x_1^2+15x_2^2$$

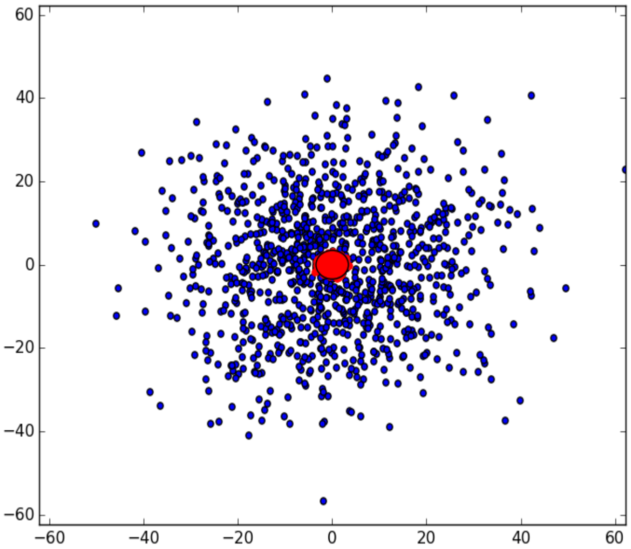
$$\frac{32\ln(256)}{15}>x_1^2+x_2^2$$

so we choose

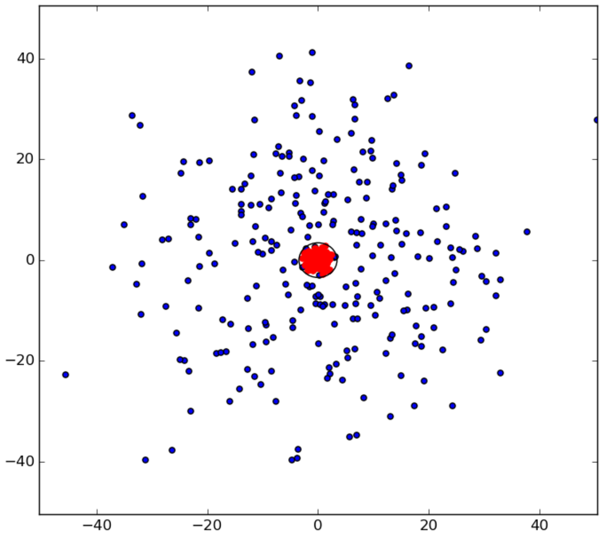
y=0 when $x_1^2+x_2^2<\frac{32\ln(256)}{15}$ otherwise we choose y=1. Computed error rate for this boundary circle with 10 000 samples was 0.0129.

Programming:

a)&b)

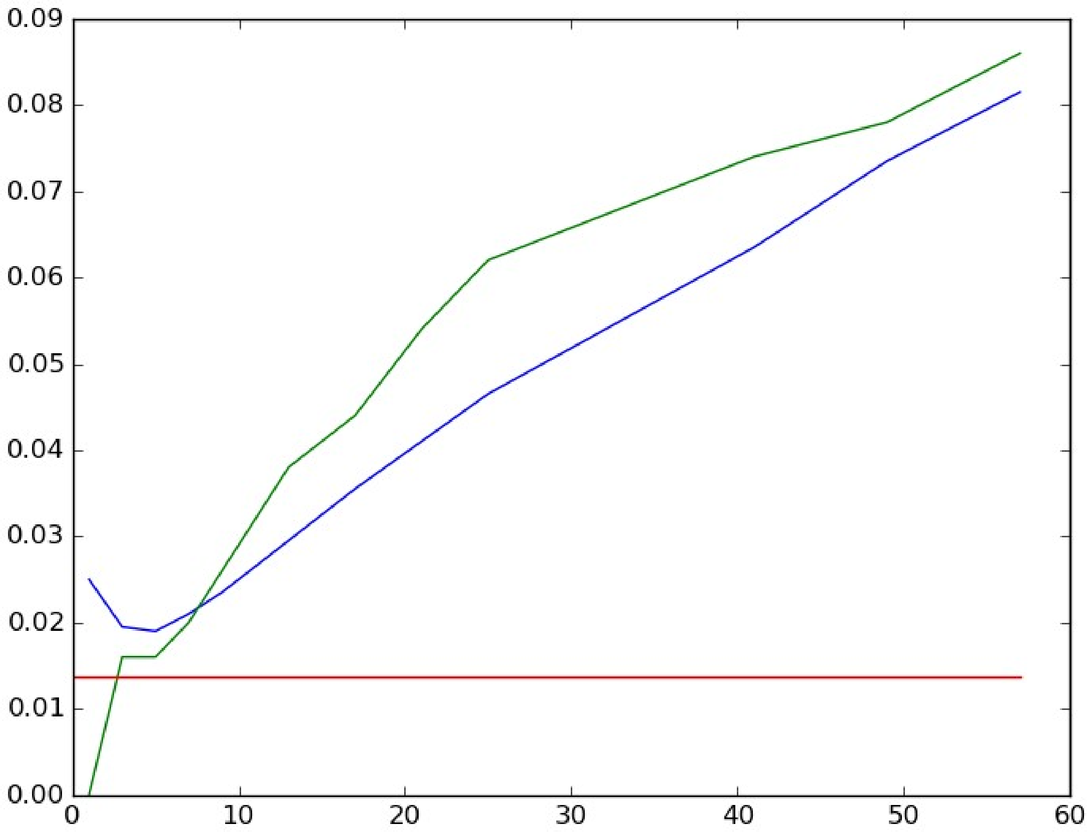


Test set with 2000 points decision boundary in black circle



Training set with 500 points decision boundary in black circle

c)



Bayes error (red), kNN on training data(green) and kNN on test data(blue) with percentage of misclassifications on y axis and k on x axis