Machine Learning

Exercise Sheet on Evaluation of Hypotheses

1. Given:

   Error rate = 6.67% and Test data size, n = 45

   For the test sample we have

   θ = Error rate = 0.0667 and $\sigma = \sqrt{\theta(1-\theta)} = 0.2495$

   According to the Central Limit Theorem, Error rate is a random variable having the Gaussian Distribution with

   Mean, $\theta_x = \theta = 0.0667$

   Standard Deviation, $\sigma_x = \frac{\sigma}{\sqrt{n}} = \frac{0.2494}{\sqrt{45}} = 0.0372$

   The 95% confidence lies between $[\theta_x - 1.96\sigma_x , \theta_x + 1.96\sigma_x]$

   Thus, ***95% confidence interval for true error*** *is*

   [-0.0062, 0.1396]

   This means that the learnt hypothesis *h1* will give an error between -0.62% to 13.96% for 95 out of 100 samples.

2. The error rate and accuracy rate of all the three hypothesis *h1, h2, h3* are given in the tabular form as

Hypothesis	**Error Rate**	**Accuracy Rate (θ_i)**	**Variance (σ_i^2) $= \frac{\theta_i(1-\theta_i)}{n}$**
h1	0.0667	0.9333	0.00138
h2	0.0889	0.9111	0.00179
h3	0.133	0.867	0.00256

i) Comparing the performance of hypothesis *h1* and *h2* on the underlying population

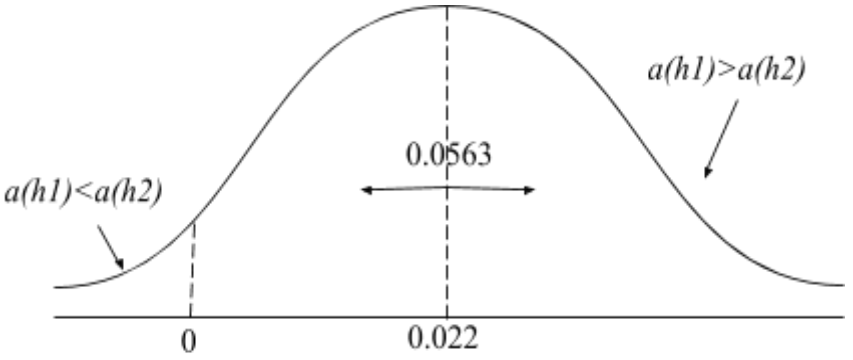
Difference of accuracy between *h1* and *h2,* $\Delta_{\theta_1-\theta_2} = 0.9333 - 0.9111 = 0.0222$

The difference of two random variables is again a random variable having gaussian distribution with

Mean, $\mu = \Delta_{\theta_1-\theta_2} = 0.022$

Standard Deviation, $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{0.00138 + 0.00179} = 0.0563$

The above scenario can be represented as



Finding z-score, for which the $\mu + z\sigma = 0$ i.e boundary upto where accuracy of hypothesis *h1* is greater than the hypothesis *h2*. So,

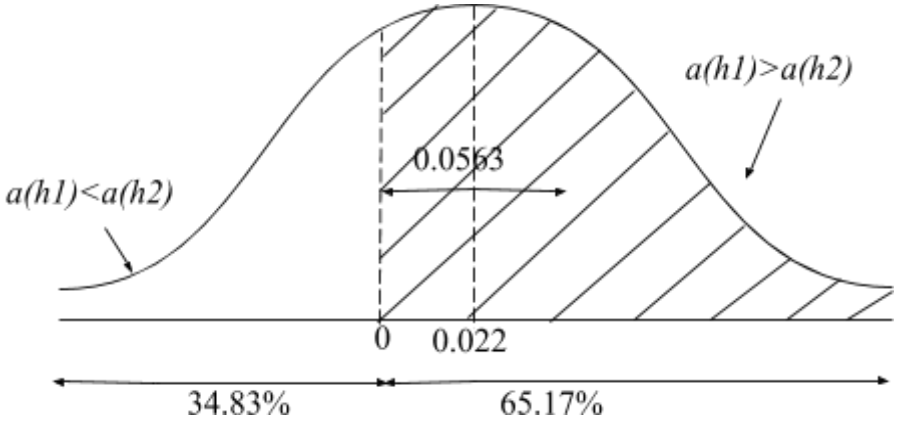
$$\mu - z\sigma = 0$$

$$0.022 - z * 0.0563 = 0$$

$$\Rightarrow z = 0.39$$

65.17% of the area of the curve comes under z = 0.39

The *test* reveals the following



Thus we can say that we are about ***65% confident that the hypothesis h1 will perform better than the hypothesis h2***

or

We can say that we are about ***65% confident that h2 will perform worse than h1 on the underlying population.***

ii) Comparing the performance of hypothesis *h1* and *h3* on the underlying population

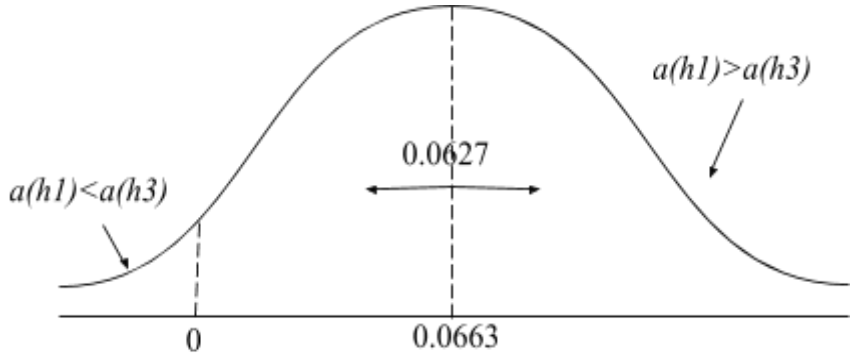
Difference of accuracy between *h1* and *h3,* $\Delta_{\theta_1-\theta_3} = 0.9333 - 0.867 = 0.0663$

The difference of two random variables is again a random variable having gaussian distribution with

Mean, $\mu = \Delta_{\theta_1-\theta_3} = 0.0663$

Standard Deviation, $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2} = \sqrt{0.00138 + 0.00256} = 0.0627$

The above scenario can be represented as



Finding z-score, for which the $\mu + z\sigma = 0$ i.e boundary upto where accuracy of hypothesis *h1* is greater than the hypothesis *h3*. So,

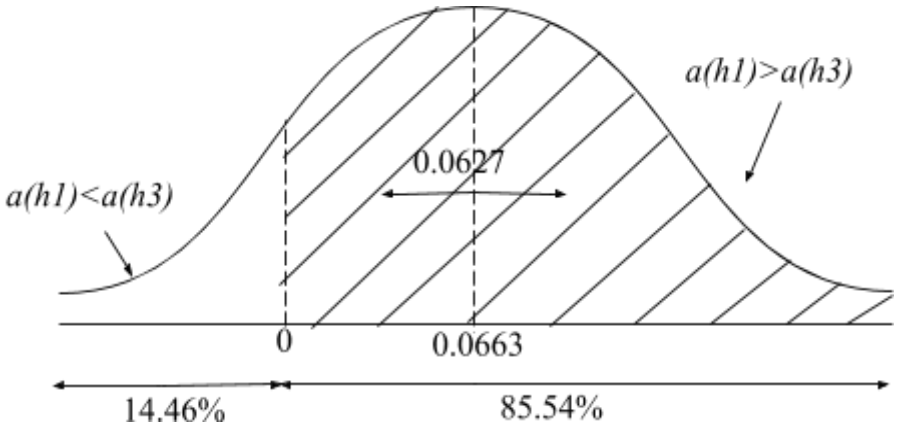
$$\mu - z\sigma = 0$$

$$0.0663 - z * 0.0627 = 0$$

$$\Rightarrow z = 1.06$$

85.54% of the area of the curve comes under z = 1.06

The *test* reveals the following



Thus we can say that we are about ***86% confident that the hypothesis h3 will perform worse than h1 on the underlying population.***

Our confidence is thus ***higher*** *for h3 to perform worse than h1 on the underlying population than we had for h2 to perform worse than h1 on the underlying population.*

3. Comparing two algorithms named as A1 and A2 by applying the 10 fold cross validation (*k = 10)* is as followed

	Accuracy Rates		
Cross Validation Fold(k)	**Favourite Algorithm(A1)**	**Decision Tree Induction(A2)**	**Y a(A1) - a(A2)**
1	0.9111	0.907	0.0041
2	0.9048	0.9052	-0.0004
3	0.9187	0.9088	0.0099
4	0.9052	0.9087	-0.0035
5	0.8988	0.9002	-0.0014
6	0.8977	0.8899	0.0078
7	0.9144	0.9098	0.0046
8	0.9088	0.9144	-0.0056
9	0.9077	0.9077	0
10	0.9089	0.9092	-0.0003
		$\Sigma a(A1) - a(A2)$	**0.0152**

The difference of two random variables is again a random variable having gaussian distribution with

Mean, $\mu = \frac{\Sigma a(A1) - a(A2)}{k} = 0.00152$

Standard deviation, σ calculated as

Y	**Y - μ**
0.0041	0.0000066564
-0.0004	0.0000036864
0.0099	0.0000702244

-0.0035	0.0000252004
-0.0014	0.0000085264
0.0078	0.0000394384
0.0046	0.0000094864
-0.0056	0.0000506944
0	0.0000023104
-0.0003	0.0000033124
$\Sigma(Y-\mu)^2$	**0.000219536**

Thus, $\sigma = \sqrt{\frac{\Sigma(Y-\mu)^2}{k}} = 0.00468$

Finding t-score, for which the $\mu + t\sigma = 0$ i.e boundary upto where *algorithm A1* will outperform *algorithm A2* . So,

$$\mu - t\sigma = 0$$

$$0.00152 - t * 0.00468 = 0$$

$$t = 0.3247$$

The degree of freedom for the given sample, *df* = k - 1 = 9

62.36% of the area of the curve comes under t = 1.06 and *df* = 9

Thus we can say that we are about ***62% confident that our favourite concept description algorithm (A1) will out-perform the decision tree induction algorithm (A2).***

4. In order to investigate the errors in classification by hypothesis *h1, h2 and h3* using *ROC* plot we need to find the *True Positive Rate(TPR) and False Positive Rate(FPR)* from the confusion matrix.

   TPR and FPR are given as

   $$\text{TPR} = \frac{\textit{No. of positive instances predicted as postive}}{\textit{Total no. of actual positive instances in the test data}}$$

   $$\text{FPR} = \frac{\textit{No. of negative instances predicted as postive}}{\textit{Total no. of actual negative instances in the test data}}$$

   For Hypothesis h1

h1		**Actual**		
		Positive	*Negative*	**Marginal Sum**
Predicted	*Positive*	29	1	30
	Negative	2	13	15
	Marginal Sum	31	14	**45**

   $$\text{TPR} = \frac{29}{29+2} = 0.935$$

   $$\text{FPR} = \frac{1}{13+1} = 0.071$$

   For Hypothesis h2

h2		**Actual**		
		Positive	*Negative*	**Marginal Sum**
Predicted	*Positive*	29	3	32
	Negative	1	12	13
	Marginal Sum	30	15	**45**

TPR = $\frac{29}{29+1} = 0.96$

FPR = $\frac{3}{3+12} = 0.2$

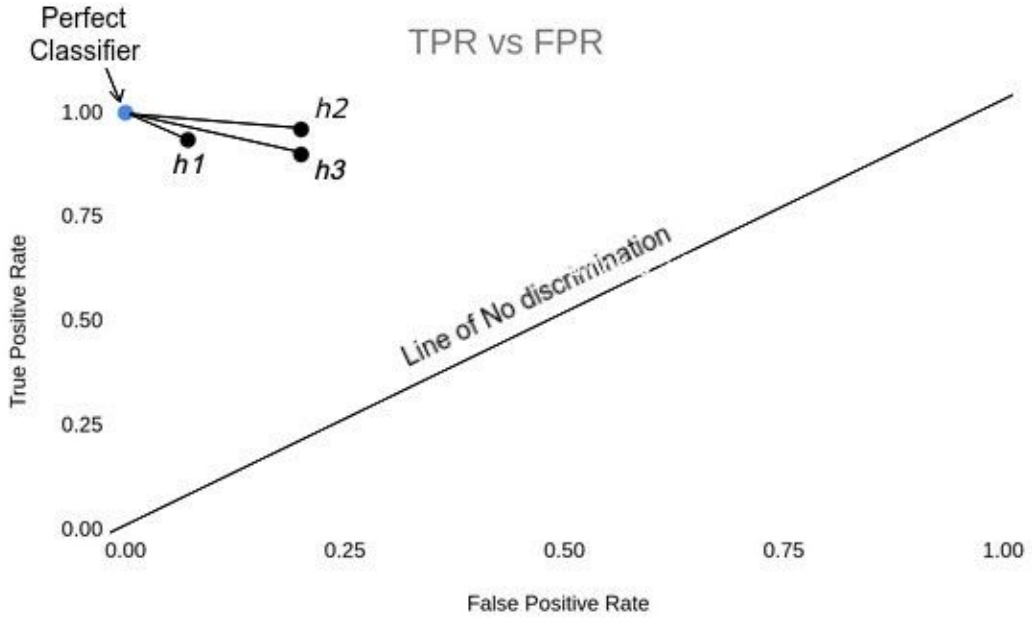
For Hypothesis h3

h2		**Actual**		
		Positive	*Negative*	**Marginal Sum**
Predicted	*Positive*	27	3	30
	Negative	3	12	15
	Marginal Sum	30	15	**45**

TPR = $\frac{27}{27+3} = 0.9$

FPR = $\frac{3}{3+12} = 0.2$

The Graphical Representation of the hypothesis *h1, h2* and *h3 is as*



i) For equal cost for false positive and false negative

Euclidean Distance = $\sqrt{w * (1 - TP)^2 + (1 - w) * FP^2}$

and AC_d = 1 - Euclidean Distance

As both false positive and false negative have equal cost thus *w = 0.5*

Hypothesis	**TPR**	**FPR**	**Euclidean Distance**	AC_d
h1	0.935	0.071	0.068	0.931
h2	0.96	0.2	0.144	0.856
h3	0.9	0.2	0.158	0.842

Larger the AC_d better is the algorithm, thus from the above case ***hypothesis h1 comes out be the best classifier of all three.***

ii) For False positives cost 4 times as much as False negative

Considering the case that False positives cost 4 times as much as False negative, we will choose **w = 0.2** and the *euclidean distance* then will be

$$\sqrt{0.2 * (1 - TP)^2 + 0.8 * FP^2}$$

Hypothesis	**TPR**	**FPR**	**Euclidean Distance**	AC_d
h1	0.935	0.071	0.0698	0.9302
h2	0.96	0.2	0.1798	0.8202
h3	0.9	0.2	0.1844	0.8156

As AC_d is again larger for hypothesis *h1,* ***hypothesis h1 comes out to be the best classifier of all three.***