CSC411 Fall 2018: Homework 1

Niloufar Afsariardchi

March 9, 2020

1 Problem 1

a) X and Y are independent with the same uniform distribution, therefore:

$$\begin{aligned} \mu &\equiv E(Y) = E(X) && (1)\\ &= \int_0^1 xf(x)dx && (2)\\ &= \int_0^1 xdx && (3)\\ &= 0.5x^2\lfloor_0^1 = \frac{1}{2} && (4) \end{aligned}$$

Similarly, we can compute the variance of X and Y:

$$\begin{aligned} \sigma^2 &\equiv Var(X) = Var(Y) && (5)\\ &= E(X^2) - \mu^2 && (6)\\ &= \int_0^1 x^2dx - \mu^2 && (7)\\ &= \frac{1}{3}x^3\lfloor_0^1 - \frac{1}{4} = \frac{1}{12}, && (8) \end{aligned}$$

where Equation 6 is simply derived in the class. Now for random variable Z, we find,

$$\begin{aligned} E(Z) &= E((X-Y)^2) && (9)\\ &= E(X^2) + E(Y^2) - 2E(XY) && (10)\\ &= 2E(X^2) - 2E(X)^2 && (11)\\ &= 2(\sigma^2 + \mu^2) - 2\mu^2 = 2\sigma^2 = \frac{1}{6} && (12) \end{aligned}$$

Equation 11 is the result of X and Y being independent. Similarly,

$$\begin{aligned} Var(Z) &= E(Z^2) - E(Z)^2 && (13)\\ &= E(X^4 + Y^4 + 6X^2Y^2 - 4X^3Y - 4Y^3X) - \frac{1}{36} && (14)\\ &= 2E(X^4) + 6E(X^2)^2 - 8E(X^3)E(Y) - \frac{1}{36} && (15)\\ &= 2E(X^4) + 6\sigma^4 + 6\mu^4 + 6\sigma^2\mu^2 - 8\mu E(X^3) - \frac{1}{36} && (16) \end{aligned}$$

Third and forth moments of X are

$$E(X^3) = \int_0^1 x^3 dx \quad (17)$$
$$= \frac{1}{4}x^4[_0^1 = \frac{1}{4} \quad (18)$$
$$E(X^4) = \int_0^1 x^4 dx \quad (19)$$
$$= \frac{1}{5}x^5[_0^1 = \frac{1}{5} \quad (20)$$

Inserting above in Equation 16, we obtain

$$Var(Z^2) = \quad (21)$$
$$2E(X^4) + 6\sigma^4 + 6\mu^4 + 6\sigma^2\mu^2 - 8\mu E(X^3) - \frac{1}{36} = \quad (22)$$
$$2/5 + 1/24 + 6/16 + 1/4 - 1 - 1/36 \simeq 0.039 \quad (23)$$

b) The expectation of squared euclidean distance is $R = \Sigma_{i=1}^d Z_i$, therefore,

$$E(R) = \Sigma_{i=1}^d E(Z_i) \quad (24)$$
$$= d \times E(Z_i) \quad (25)$$
$$= d \times E(Z) \quad (26)$$

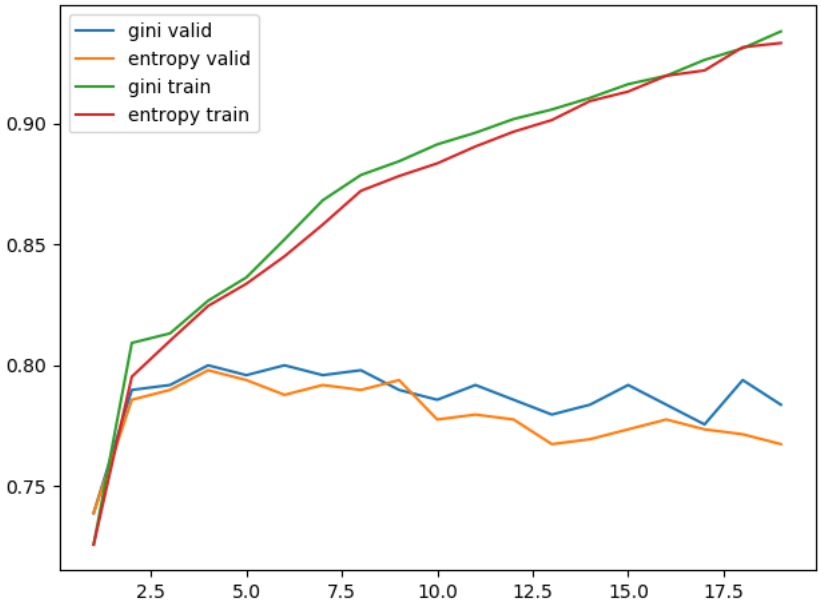
Note that Equation 25 is the result of the Z_is being independent. Since X_is and Y_is are independent, Z_is are also independent and identically distributed.

$$Var(R) = E(R^2) - E(R)^2 \quad (27)$$
$$= \sum_j^d \sum_i^d E(Z_i Z_j) - E(R)^2 \quad (28)$$
$$= d(Var(Z) + E(Z)^2) + d(d-1)E(Z)^2 - E(R)^2 \quad (29)$$
$$= d \times Var(Z) + d^2 E(Z)^2 - d^2 E(Z)^2 \quad (30)$$
$$= d \times Var(Z) \quad (31)$$

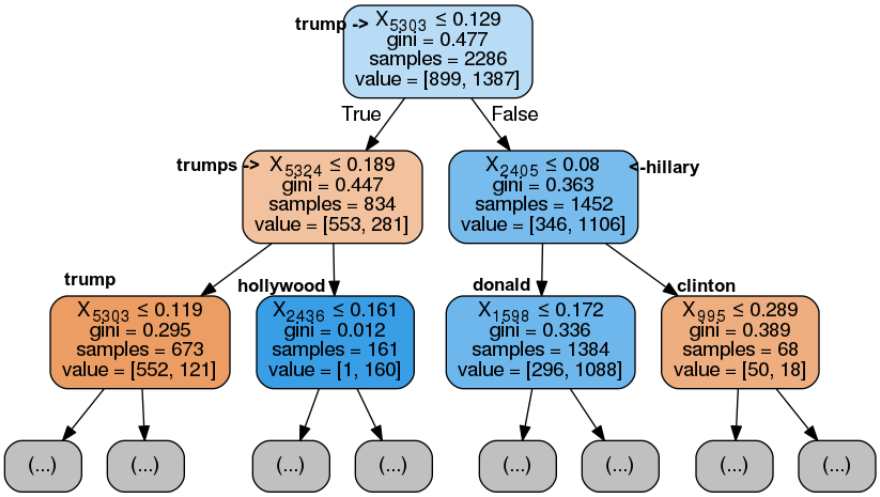
2 Problem 2

b) The output is:
Gini maxdepth= 1 — training: 0.725721784777 validation: 0.738775510204
Entropy maxdepth= 1 — training: 0.725721784777 validation: 0.738775510204
Gini maxdepth= 2 — training: 0.80927384077 validation: 0.789795918367
Entropy maxdepth= 2 — training: 0.795275590551 validation: 0.785714285714
Gini maxdepth= 3 — training: 0.813210848644 validation: 0.791836734694
Entropy maxdepth= 3 — training: 0.810148731409 validation: 0.789795918367
Gini maxdepth= 4 — training: 0.826771653543 validation: 0.8
Entropy maxdepth= 4 — training: 0.824584426947 validation: 0.797959183673
Gini maxdepth= 5 — training: 0.836395450569 validation: 0.795918367347
Entropy maxdepth= 5 — training: 0.833770778653 validation: 0.79387755102

Gini maxdepth= 6 — training: 0.852143482065 validation: 0.8
Entropy maxdepth= 6 — training: 0.845144356955 validation: 0.787755102041
Gini maxdepth= 7 — training: 0.86832895888 validation: 0.795918367347
Entropy maxdepth= 7 — training: 0.858267716535 validation: 0.791836734694
Gini maxdepth= 8 — training: 0.878827646544 validation: 0.797959183673
Entropy maxdepth= 8 — training: 0.872265966754 validation: 0.789795918367
Gini maxdepth= 9 — training: 0.884514435696 validation: 0.789795918367
Entropy maxdepth= 9 — training: 0.878390201225 validation: 0.79387755102
Gini maxdepth= 10 — training: 0.891513560805 validation: 0.785714285714
Entropy maxdepth= 10 — training: 0.883639545057 validation: 0.777551020408
Gini maxdepth= 11 — training: 0.896325459318 validation: 0.791836734694
Entropy maxdepth= 11 — training: 0.890638670166 validation: 0.779591836735
Gini maxdepth= 12 — training: 0.902012248469 validation: 0.785714285714
Entropy maxdepth= 12 — training: 0.896762904637 validation: 0.777551020408
Gini maxdepth= 13 — training: 0.905949256343 validation: 0.779591836735
Entropy maxdepth= 13 — training: 0.90157480315 validation: 0.767346938776
Gini maxdepth= 14 — training: 0.910761154856 validation: 0.783673469388
Entropy maxdepth= 14 — training: 0.909448818898 validation: 0.769387755102
Gini maxdepth= 15 — training: 0.916447944007 validation: 0.791836734694
Entropy maxdepth= 15 — training: 0.913385826772 validation: 0.773469387755
Gini maxdepth= 16 — training: 0.919947506562 validation: 0.783673469388
Entropy maxdepth= 16 — training: 0.919947506562 validation: 0.777551020408
Gini maxdepth= 17 — training: 0.926509186352 validation: 0.775510204082
Entropy maxdepth= 17 — training: 0.922134733158 validation: 0.773469387755
Gini maxdepth= 18 — training: 0.931321084864 validation: 0.79387755102
Entropy maxdepth= 18 — training: 0.931758530184 validation: 0.771428571429
Gini maxdepth= 19 — training: 0.938320209974 validation: 0.783673469388
Entropy maxdepth= 19 — training: 0.933508311461 validation: 0.767346938776
Here is a plot of the output:



c) Based on the plot above, it seems that there is no significant accuracy gain on the validation set after max depth of 6. Also, Gini criterion slightly outperforms entropy criterion. Below is the tree classifier with best hyper-parameters. Note that I used scikit-learn's TfidfVectorizer for building the vocabulary and tokenizing the input text.



d) I calculate IG for several split including "trump","clinton","hillary","donald". Since I used TfidfVectorizer, I needed my vocabulary list to map words to feature indices as well as a threshold on the score of the split. For all cases I chose the threshold in my top split, 0.129. Here is the IGs of different splits:

trump 0.127119979504

clinton 0.0085884191235

hillary 0.0377142497606

donald 0.05878904984

Clearly, the split on "trump" word gives us the highest IG.