145 Introduction to Data Mining HW1

Calvin Chen

004943505

1.

1.1

(a)

$$X = \begin{bmatrix} 1 & 60 \\ 1 & 70 \\ 1 & 62 \\ 1 & 72 \\ 1 & 65 \end{bmatrix}, \quad y = \begin{bmatrix} 130 \\ 155 \\ 125 \\ 162 \\ 150 \end{bmatrix}$$

$$\hat{\beta} = \left(\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 60 & 70 & 62 & 72 & 65 \end{bmatrix} \begin{bmatrix} 1 & 60 \\ 1 & 70 \\ 1 & 62 \\ 1 & 72 \\ 1 & 65 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 60 & 70 & 62 & 72 & 65 \end{bmatrix} \begin{bmatrix} 130 \\ 155 \\ 125 \\ 162 \\ 150 \end{bmatrix}$$

$$= \left(\begin{bmatrix} 5 & 329 \\ 329 & 21753 \end{bmatrix} \right)^{-1} \begin{bmatrix} 722 \\ 47814 \end{bmatrix}$$

$$= \frac{1}{524} \begin{bmatrix} 21753 & -329 \\ -329 & 5 \end{bmatrix} \begin{bmatrix} 722 \\ 47814 \end{bmatrix} = \frac{1}{524} \begin{bmatrix} -25140 \\ 1532 \end{bmatrix} = \begin{bmatrix} -47.98 \\ 2.92 \end{bmatrix}$$

(b)

Student 1: $-47.98 + 2.92 * 60 = 127.22$

Student 2: $-47.98 + 2.92 * 70 = 156.42$

Student 3: $-47.98 + 2.92 * 62 = 133.06$

Student 4: $-47.98 + 2.92 * 72 = 162.26$

Student 5: $-47.98 + 2.92 * 65 = 141.82$

1.2

(a)

The weights and MSE in different version are not the same.

For batch gradient descent, it run through all the test data to update the weight; whereas stochastic gradient descent uses only one data in each update of the weights. And close form uses least squares to estimate the weights.

```
Closed Form Without Normalization
Beta:
[-0.0862246   0.05340575  0.65803045  0.41731923 -0.01772481  0.30069864
  1.02871152  0.48383363  0.26685697  0.04573456  0.31944742  1.14776959
  0.29366213  0.41491543  0.85180482 -0.05950309  0.47235562  0.46198106
  0.00497427  0.0205398   0.41310473  0.98508025  0.15573467  0.8618602
  0.41974331 -0.06893699  0.33317496  0.27766637 -0.04184791 -0.23599504
  0.15020297  0.37745027  0.80256455  0.16053288  0.2744667   0.63461071
  0.74135259  0.56079776  0.94058723 -0.0432542   0.80803615  0.93967722
  0.12225161 -0.19933624  0.09398732  0.11412993  0.35479619  0.78582876
  0.38900433  0.11804526  0.67618837  0.70377377  0.05526258 -0.24919095
  0.87339793 -0.01381723  0.83138416  0.90569236  0.39980648  0.25235308
  0.69692397 -0.00949757  0.17676599  0.45822485  0.02743899  1.16718165
  0.04176352  1.01993881  0.56015024 -0.29761224  0.3177761   0.55781578
  1.1376088   0.55190283  0.4099807   0.91987238  1.34076835  0.53297825
  0.63648277  0.22140583  0.21469531 -0.00609269  0.82898663  0.46891532
 -0.25571565  0.1972989   1.38639797  0.87219453  0.65782257  0.54983464
  1.11698567  0.94267463  0.79030138  0.30055848  0.53288973  0.22873689
  0.86702876  0.98591924  0.08132528  0.30834368  0.70121488]

MSE:  4.39609786082
```

```
Batch Gradient Without Normalization
Step:  277993
Beta:
[-0.08591641  0.05340513  0.65802932  0.41731352 -0.01772567  0.3006924
  1.02870368  0.48382507  0.26685263  0.04572647  0.31944062  1.14776048
  0.29365535  0.41490666  0.85179876 -0.0595114   0.47235019  0.46197245
  0.00496603  0.02053546  0.41309652  0.98507556  0.15572541  0.86185487
  0.41973916 -0.06894328  0.33316928  0.27766077 -0.04186029 -0.23600335
  0.15020121  0.37744187  0.8025563   0.16052926  0.27445941  0.63460247
  0.74134744  0.56079149  0.94057699 -0.04325966  0.808029    0.93967068
  0.12224922 -0.1993428   0.09398035  0.11412415  0.35479127  0.78582269
  0.38900229  0.11803836  0.67618212  0.70376774  0.05525309 -0.2491959
  0.87339345 -0.01382283  0.831377    0.90568597  0.39980006  0.25234819
  0.69692076 -0.00950408  0.17676184  0.45821889  0.02742915  1.16717321
  0.04175878  1.01993274  0.56014129 -0.2976159   0.31777218  0.55781012
  1.13760164  0.55189329  0.40998191  0.91986805  1.34076065  0.53297409
  0.63647711  0.22140133  0.21469042 -0.00609686  0.82898029  0.46890736
 -0.25572296  0.19729072  1.38639215  0.87219305  0.65781697  0.54982704
  1.11697763  0.94267076  0.79029359  0.30055174  0.53288539  0.22872743
  0.86701975  0.9859145   0.08131992  0.30833799  0.70121206]

MSE:  4.39609852466
```

```
Stochastic Gradient Without Normalization
Step:  5814
Beta:
[ 0.4726906   0.05252962  0.64584253  0.40187961  0.00655931  0.27827303
  0.97741144  0.47336841  0.27911165  0.01510089  0.31676338  1.14055254
  0.29710437  0.42036244  0.83244084 -0.0697041   0.46213277  0.45598868
  0.00256003  0.01914751  0.3954126   0.97570473  0.17281891  0.8369735
  0.40393156 -0.0557204   0.30762504  0.27630468 -0.04585385 -0.22989501
  0.14569454  0.36856827  0.78167209  0.16406548  0.26617355  0.61670469
  0.72055276  0.54321041  0.90724699 -0.05342281  0.79624507  0.92449566
  0.1105473  -0.20634034  0.09095631  0.12348402  0.33239431  0.76721219
  0.37700494  0.12477562  0.67293193  0.67399185  0.03669845 -0.25305749
  0.88057572 -0.01632913  0.8036066   0.88862355  0.40320974  0.2427044
  0.67651919 -0.03365885  0.16177904  0.44546403  0.02699722  1.14551893
  0.05128868  1.00746896  0.54278907 -0.29181164  0.31558196  0.57125914
  1.11900463  0.52128047  0.40782474  0.90129886  1.30634835  0.50391946
  0.60945143  0.20134875  0.18469516 -0.01053448  0.81737303  0.46610183
 -0.24267096  0.18646509  1.3757151   0.85885187  0.65970738  0.55781428
  1.0817365   0.94034977  0.76221445  0.29589653  0.52214676  0.20476847
  0.83698012  0.95279395  0.06289908  0.30708873  0.67183316]

MSE:  4.38404109087
```

(b)

After normalization, all the weights and MSE of the three versions change and can eliminate the difference between different feature because of normalizing them into same scale.

```
Closed Form With Normalization
Beta:
[  2.27729720e+01   1.53267685e-01   1.85400036e-01   1.20001101e-01
  -5.02894960e-03   8.91855522e-02   2.85477509e-01   1.40249729e-01
   7.58001703e-02   1.29653087e-02   9.40114997e-02   3.31501951e-01
   8.48405150e-02   1.19998020e-01   2.42101087e-01  -1.70904428e-02
   1.37119556e-01   1.35350218e-01   1.41619004e-03   5.96423043e-03
   1.15830867e-01   2.84837752e-01   4.40248244e-02   2.49185633e-01
   1.20285952e-01  -1.97966211e-02   9.78939759e-02   8.05403060e-02
  -1.21241111e-02  -6.77059821e-02   4.42940642e-02   1.07814670e-01
   2.27982170e-01   4.72154203e-02   7.98729034e-02   1.82957097e-01
   2.10609705e-01   1.62079663e-01   2.74455584e-01  -1.24456123e-02
   2.32346197e-01   2.68821067e-01   3.49745502e-02  -5.73174263e-02
   2.74558199e-02   3.22366923e-02   1.03219840e-01   2.23792899e-01
   1.12445398e-01   3.34223468e-02   1.96611852e-01   2.04171370e-01
   1.61259528e-02  -7.12316220e-02   2.51757075e-01  -3.88735810e-03
   2.31055679e-01   2.65481860e-01   1.14239087e-01   7.19519080e-02
   2.03225977e-01  -2.77922653e-03   5.10043840e-02   1.31478537e-01
   7.74623329e-03   3.36781203e-01   1.19518825e-02   2.98145298e-01
   1.64253970e-01  -8.57326109e-02   9.04810592e-02   1.57878654e-01
   3.30578812e-01   1.58142457e-01   1.17519641e-01   2.66603450e-01
   3.90619100e-01   1.54573813e-01   1.82230684e-01   6.25165215e-02
   6.11873098e-02  -1.74345404e-03   2.34361003e-01   1.35158424e-01
  -7.34879378e-02   5.72871764e-02   4.02966409e-01   2.50642329e-01
   1.87572968e-01   1.57855445e-01   3.18225717e-01   2.66144412e-01
   2.29911686e-01   8.53225833e-02   1.56706806e-01   6.57087894e-02
   2.52781623e-01   2.90068806e-01   2.33284776e-02   9.01229905e-02
   2.03998476e-01]

MSE:  4.40454594906
```

```
Batch Gradient With Normalization
Step:  2882
Beta:
[  2.27729720e+01   1.53267643e-01   1.85399940e-01   1.20000994e-01
  -5.02900488e-03   8.91854493e-02   2.85477555e-01   1.40249879e-01
   7.58002772e-02   1.29652510e-02   9.40114970e-02   3.31501978e-01
   8.48406412e-02   1.19998261e-01   2.42101085e-01  -1.70903535e-02
   1.37119527e-01   1.35350299e-01   1.41627211e-03   5.96415447e-03
   1.15830908e-01   2.84837863e-01   4.40249905e-02   2.49185533e-01
   1.20285904e-01  -1.97965304e-02   9.78939965e-02   8.05403730e-02
  -1.21238416e-02  -6.77058144e-02   4.42939488e-02   1.07814862e-01
   2.27982240e-01   4.72153902e-02   7.98729252e-02   1.82957108e-01
   2.10609670e-01   1.62079672e-01   2.74455647e-01  -1.24455447e-02
   2.32346291e-01   2.68821034e-01   3.49744819e-02  -5.73173997e-02
   2.74557824e-02   3.22368299e-02   1.03219861e-01   2.23792870e-01
   1.12445361e-01   3.34223860e-02   1.96611901e-01   2.04171259e-01
   1.61259881e-02  -7.12315815e-02   2.51757074e-01  -3.88737096e-03
   2.31055798e-01   2.65481993e-01   1.14239091e-01   7.19519378e-02
   2.03225951e-01  -2.77921048e-03   5.10043321e-02   1.31478625e-01
   7.74633978e-03   3.36781256e-01   1.19519427e-02   2.98145387e-01
   1.64254089e-01  -8.57325789e-02   9.04811182e-02   1.57878661e-01
   3.30578922e-01   1.58142509e-01   1.17519517e-01   2.66603433e-01
   3.90619125e-01   1.54573835e-01   1.82230654e-01   6.25164842e-02
   6.11872499e-02  -1.74345246e-03   2.34361121e-01   1.35158557e-01
  -7.34879950e-02   5.72871796e-02   4.02966420e-01   2.50642185e-01
   1.87573016e-01   1.57855528e-01   3.18225691e-01   2.66144388e-01
   2.29911654e-01   8.53225347e-02   1.56706745e-01   6.57088564e-02
   2.52781562e-01   2.90068636e-01   2.33284819e-02   9.01229765e-02
   2.03998357e-01]

MSE:  4.40454583846
```

```
Stochastic Gradient With Normalization
Step:  944
Beta:
[  2.27713098e+01   1.54422779e-01   1.84689157e-01   1.19346560e-01
  -6.30295190e-03   8.90611974e-02   2.86492679e-01   1.43403523e-01
   7.61551379e-02   1.36301991e-02   9.41347164e-02   3.32652375e-01
   8.63753310e-02   1.22705230e-01   2.42021487e-01  -1.49796130e-02
   1.36078440e-01   1.36792890e-01   1.42003298e-03   5.59255868e-03
   1.15948815e-01   2.85447149e-01   4.60597066e-02   2.48384701e-01
   1.20285248e-01  -1.89808547e-02   9.84546098e-02   8.12315338e-02
  -9.01061721e-03  -6.48366519e-02   4.16797385e-02   1.09695183e-01
   2.29373345e-01   4.67398770e-02   8.00838062e-02   1.83929606e-01
   2.11361537e-01   1.61935179e-01   2.75823214e-01  -1.03896334e-02
   2.33358648e-01   2.68654668e-01   3.51837767e-02  -5.62650729e-02
   2.83252552e-02   3.37546810e-02   1.03135621e-01   2.23648338e-01
   1.11136622e-01   3.30442986e-02   1.97173590e-01   2.02498834e-01
   1.87996782e-02  -7.08247398e-02   2.51489105e-01  -3.21045303e-03
   2.32059421e-01   2.66983801e-01   1.15119753e-01   7.26942950e-02
   2.03026043e-01  -1.57368113e-03   4.99582600e-02   1.31994035e-01
   9.29496556e-03   3.37323711e-01   1.25410832e-02   2.99220983e-01
   1.65213002e-01  -8.56557796e-02   9.00236361e-02   1.58131260e-01
   3.32498597e-01   1.59626661e-01   1.16068291e-01   2.65640952e-01
   3.91591139e-01   1.54151697e-01   1.82692862e-01   6.11923093e-02
   6.08272241e-02  -9.24409619e-04   2.35249711e-01   1.37320344e-01
  -7.41572081e-02   5.81093988e-02   4.03279564e-01   2.49363512e-01
   1.87812138e-01   1.60170840e-01   3.18039078e-01   2.65982271e-01
   2.30267209e-01   8.42330850e-02   1.55384947e-01   6.74336026e-02
   2.53416962e-01   2.88197120e-01   2.41872703e-02   9.05289515e-02
   2.02436826e-01]

MSE:  4.40155749744
```

2.

(a)

$$x_i = \begin{bmatrix} x_{i0} \\ x_{i1} \\ x_{i2} \end{bmatrix}, \quad \beta_i = \begin{bmatrix} \beta_{i0} \\ \beta_{i1} \\ \beta_{i2} \end{bmatrix}, \quad y_i$$

$x_{00} = x_{10} = x_{20} = 1$

$x_{01} = 60 \quad x_{11} = 64 \quad x_{21} = 73$

$x_{02} = 155 \quad x_{12} = 135 \quad x_{22} = 170$

$y_0 = 0 \quad y_1 = 1 \quad y_2 = 1$

$$\begin{aligned} L = {} & y_0(x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}) - \log(1 + e^{x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}}) \\ & + y_1(x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}) - \log(1 + e^{x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}}) \\ & + y_2(x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}) - \log(1 + e^{x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}}) \end{aligned}$$

(b)

$$\begin{aligned} \frac{\partial L(\beta)}{\beta_0} = {} & x_{00}\left(y_0 - \frac{e^{x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}}}{1 + e^{x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}}}\right) + x_{01}\left(y_0 - \frac{e^{x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}}}{1 + e^{x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}}}\right) \\ & + x_{02}\left(y_0 - \frac{e^{x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}}}{1 + e^{x_{00}\beta_{00} + x_{01}\beta_{01} + x_{02}\beta_{02}}}\right) \end{aligned}$$

$$\begin{aligned} \frac{\partial L(\beta)}{\beta_1} = {} & x_{10}\left(y_1 - \frac{e^{x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}}}{1 + e^{x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}}}\right) + x_{11}\left(y_1 - \frac{e^{x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}}}{1 + e^{x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}}}\right) \\ & + x_{12}\left(y_1 - \frac{e^{x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}}}{1 + e^{x_{10}\beta_{10} + x_{11}\beta_{11} + x_{12}\beta_{12}}}\right) \end{aligned}$$

$$\begin{aligned} \frac{\partial L(\beta)}{\beta_2} = {} & x_{20}\left(y_2 - \frac{e^{x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}}}{1 + e^{x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}}}\right) + x_{21}\left(y_2 - \frac{e^{x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}}}{1 + e^{x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}}}\right) \\ & + x_{22}\left(y_2 - \frac{e^{x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}}}{1 + e^{x_{20}\beta_{20} + x_{21}\beta_{21} + x_{22}\beta_{22}}}\right) \end{aligned}$$

(c)

$$H_{00} = -x_{00}x_{00}\frac{e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}}}{(1+e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}})^2} - x_{10}x_{10}\frac{e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}}}{(1+e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}})^2} - x_{20}x_{20}\frac{e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}}}{(1+e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}})^2}$$

$$H_{01} = H_{10} = -x_{01}x_{00}\frac{e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}}}{(1+e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}})^2} - x_{11}x_{10}\frac{e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}}}{(1+e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}})^2} - x_{21}x_{20}\frac{e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}}}{(1+e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}})^2}$$

$$H_{02} = H_{20} = -x_{02}x_{00}\frac{e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}}}{(1+e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}})^2} - x_{12}x_{10}\frac{e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}}}{(1+e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}})^2} - x_{22}x_{20}\frac{e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}}}{(1+e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}})^2}$$

$$H_{11} = -x_{01}x_{01}\frac{e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}}}{(1+e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}})^2} - x_{11}x_{11}\frac{e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}}}{(1+e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}})^2} - x_{21}x_{21}\frac{e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}}}{(1+e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}})^2}$$

$$H_{12} = H_{21} = -x_{01}x_{02}\frac{e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}}}{(1+e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}})^2} - x_{11}x_{12}\frac{e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}}}{(1+e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}})^2} - x_{21}x_{22}\frac{e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}}}{(1+e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}})^2}$$

$$H_{22} = -x_{02}x_{02}\frac{e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}}}{(1+e^{\beta_0 x_{00}+\beta_1 x_{01}+\beta_2 x_{02}})^2} - x_{12}x_{12}\frac{e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}}}{(1+e^{\beta_0 x_{10}+\beta_1 x_{11}+\beta_2 x_{12}})^2} - x_{22}x_{22}\frac{e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}}}{(1+e^{\beta_0 x_{20}+\beta_1 x_{21}+\beta_2 x_{22}})^2}$$

(d)

```
Cs-210-33:LogisticRegression calvinyhchen$ python logicsticRegression.py
normalized x
[[ 1.         -0.8510645   0.0949158 ]
 [ 1.         -0.25031309 -1.0440738 ]
 [ 1.          1.10137758  0.949158  ]]
Hessian:
[[-0.71633282  0.03616987  0.03108301]
 [ 0.03616987 -0.45966315 -0.27181836]
 [ 0.03108301 -0.27181836 -0.46981226]]
Gradients:
[[ 0.32122958  0.65905937 -0.28823923]]
iter parameter:
[0.7433165000535708, 3.0105926494161808, -1.9280726780622577]
normalized x
[[ 1.         -0.8510645   0.0949158 ]
 [ 1.         -0.25031309 -1.0440738 ]
 [ 1.          1.10137758  0.949158  ]]
Hessian:
[[-0.29734191  0.01883124  0.01617043]
 [ 0.01883124 -0.18891993 -0.11062189]
 [ 0.01617043 -0.11062189 -0.19418955]]
Gradients:
[[ 0.09707802  0.17853591 -0.04321817]]
iter2 parameter:
[1.112486410978871 9, 4.6523851139698094, -3.0551504374064309]
```

3.

3.1

$$Info(D) = I(10,10) = -\frac{10}{20}\log_2\frac{10}{20} - \frac{10}{20}\log_2\frac{10}{20} = 1$$

$$Info_{hand}(D) = \frac{8}{20} I(6,2) + \frac{12}{20} I(4,8)$$

$$= \frac{8}{20}\left(-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}\right) + \frac{12}{20}\left(-\frac{4}{12}\log_2\frac{4}{12} - \frac{8}{12}\log_2\frac{8}{12}\right)$$

$$= 0.3245112 + 0.5509775 = 0.8755$$

$$Info_{water}(D) = \frac{10}{20} I(4,6) + \frac{10}{20} I(6,4)$$

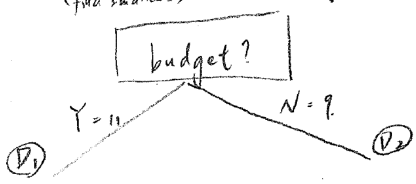
$$= I(4,6) = \left(-\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}\right) = 0.9710$$

$$Info_{budget}(D) = \frac{11}{20} I(9,2) + \frac{9}{20} I(1,8).$$

$$= \frac{11}{20}\left(-\frac{9}{11}\log_2\frac{9}{11} - \frac{2}{11}\log_2\frac{2}{11}\right) + \frac{9}{20}\left(-\frac{1}{9}\log_2\frac{1}{9} - \frac{8}{9}\log_2\frac{8}{9}\right)$$

$$= 0.376221 + 0.226466 = 0.6027.$$

$\Rightarrow Gain_x = Info(D) - \underbrace{Info_x(D)}_{\text{(find smallest)}}$ $\Rightarrow$ $Gain_{budget}$ would be the largest.

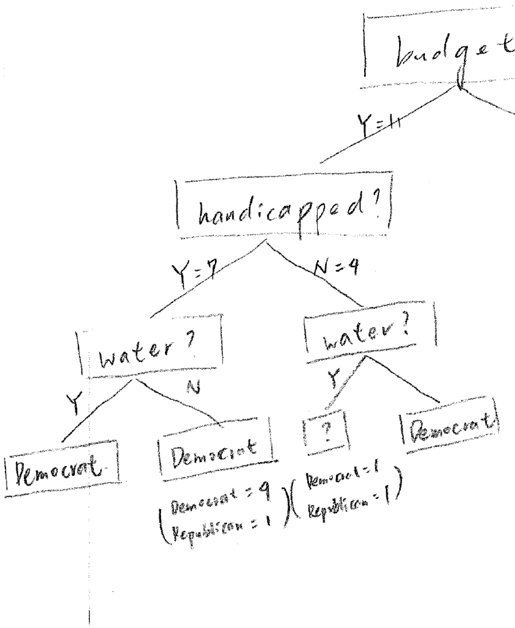


$$Info(D_1) = I(9,2) = -\frac{9}{11}\log_2\frac{9}{11} - \frac{2}{11}\log_2\frac{2}{11} = 0.6840$$

$$Info_{hand}(D_1) = \frac{7}{11} I(6,1) + \frac{4}{11} I(3,1) = \frac{7}{11}\left(-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right) + \frac{4}{11}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)$$

$$= 0.376519 + 0.295010 = 0.6715$$

$Info_{water}(D_1) = \frac{4}{11} I(3,1) + \frac{7}{11} I(6,1) = Info_{hand}(D_1)$ $\Rightarrow$ Gain would be the same.



(D_2)

$$\text{Info}(D_2) = I(1,8) = -\frac{1}{9}\log_2\frac{1}{9} - \frac{8}{9}\log_2\frac{8}{9} = 0.5033$$

$$\text{Info}_{hand}(D_2) = \cancel{\frac{1}{9}I(0,1)}^{\,0} + \frac{8}{9}I(1,7)$$

$$= \frac{8}{9}\left(-\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8}\right)$$

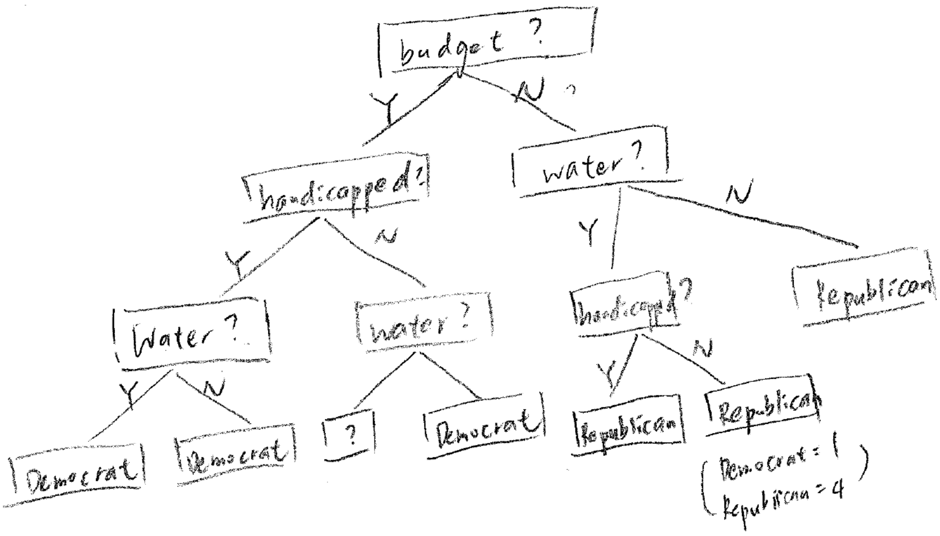
$$= 0.483168$$

$$\text{Info}_{water}(D_2) = \frac{6}{9}I(1,5) + \cancel{\frac{3}{9}I(0,3)}^{\,0}$$

$$= \frac{6}{9}\left(-\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6}\right)$$

$$= 0.4333$$

$\Rightarrow \text{Gain}_{water} > \text{Gain}(\text{handicapped})$



3.2

```
Cs-210-33:DecisionTree calvinyhchen$ python3 DecisionTree.py house-votes-84.data
Fold-1: 0.9310344827586207
Fold-2: 0.9080459770114943
Fold-3: 0.9655172413793104
Fold-4: 0.8850574712643678
Fold-5: 0.9310344827586207
5-CV Accuracy = 0.9241379310344827
```

```
Cs-210-33:DecisionTree calvinyhchen$ python3 DecisionTree.py house-votes-84.data 1
Fold-1: 0.9195402298850575
Fold-2: 0.9195402298850575
Fold-3: 0.9540229885057471
Fold-4: 0.8850574712643678
Fold-5: 0.9310344827586207
5-CV Accuracy = 0.9218390804597701
```

```
Cs-210-33:DecisionTree calvinyhchen$ python3 DecisionTree.py tic-tac-toe.data
Fold-1: 0.875
Fold-2: 0.8333333333333334
Fold-3: 0.828125
Fold-4: 0.8167539267015707
Fold-5: 0.8272251308900523
5-CV Accuracy = 0.8360874781849912
```

```
Cs-210-33:DecisionTree calvinyhchen$ python3 DecisionTree.py tic-tac-toe.data 1
Fold-1: 0.8802083333333334
Fold-2: 0.8229166666666666
Fold-3: 0.8177083333333334
Fold-4: 0.837696335078534
Fold-5: 0.837696335078534
5-CV Accuracy = 0.8392452006980804
```

With dataset house-vote, I would like to choose information gain as the attribute selection measure since it has better performance.

But for dataset tic-tac-toe, I would like to choose gain ratio as the attribute selection measure since the previous used information gain measure is biased towards attributes with a large number of values