# HOMEWORK 2

30/30

1> (a) $I(X,Y) = H(X) - H(X|Y)$

$$= -E_p(\ln p(X)) + E_p(\ln p(X|Y))$$

$$= -E_p(\ln p(X) - \ln p(X|Y))$$ [from linearity of expectations].

$$= -E_p \ln\left(\frac{p(X)}{p(X|Y)}\right)$$

$$= -E_p \ln\left(\frac{p(X)\,p(Y)}{p(X,Y)}\right)$$

(b) $H(X,Y) = -E_p(\ln p(X,Y))$

$$= -E_p(\ln p(X|Y)\,p(Y))$$

$$= -E_p(\ln p(X|Y) + \ln p(Y))$$

$$= -E_p \ln p(X|Y) - E_p \ln p(Y)$$

$$= H(X|Y) + H(Y)$$

(c) $I(X,Y|Z) = H(X|Z) - H(X|Y,Z)$

$$= -\{E_p \ln p(X|Z) - E_p \ln p(X|Y,Z)$$

$$= -\{E_p \ln p(X|Z) - E_p \ln \frac{p(X,Y,Z)}{p(Y,Z)}\}$$

$$= -\{E_p \ln p(X|Z) - E_p \ln \frac{p(X,Y,Z)\,p(Z)}{p(Y,Z)\,p(Z)}$$

$$= -\left\{ E_p \ln p(x,z) - E_p \ln \frac{p(x,y/z)}{p(y/z)} \right\}$$

$$= - E_p \ln \frac{p(x/z)\, p(y/z)}{p(x,y/z)}$$

The conditional independence assumption, that guarantees $I(x, y/z) = 0$ is:

$$p(x/z)\, p(y/z) = p(x, y/z).$$

2) (a)

$$\sigma_x^2 = E((x - Ex)^2)$$

$$\sigma_y^2 = E((y - Ey)^2)$$

From Cauchy - Schwarz inequality,

$$\left\{ E[(x - Ex)(y - Ey)] \right\}^2 \leq E[(x - Ex)^2]\, E[(y - Ey)^2]$$

$$\{cov(x, y)\}^2 \leq \sigma_x^2 \sigma_y^2$$

$$\left\{ \frac{cov(x, y)}{\sigma_x \sigma_y} \right\}^2 \leq 1$$

$$\rho_{xy}^2 \leq 1$$

$$|\rho_{xy}| \leq 1.$$

(b) $x = ay \quad \Rightarrow \quad E(x) = aE(y)$ [linearity of expectation]

$$\mathrm{cov}(X,Y) = E((X-EX)(Y-EY))$$

$$= E((X-EX)\,a(X-EX)]$$

$$= \cancel{a}\, E(a(X-EX)^2)$$

$$= a\, E[(X-EX)^2]$$

$$= a\,\sigma_x^2$$

$$\sigma_x^2 = a^2 \sigma_y^2$$

$$\sigma_x = |a|\,\sigma_y .$$

$$\cancel{\therefore \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} = 1}$$

$$\rho_{xy} = \frac{\mathrm{cov}(X,Y)}{\sigma_x \sigma_y} = \frac{a\cancel{\sigma_x^2}}{|a|\cancel{\sigma_x}\cdot\cancel{\sigma_x}} = \frac{a}{|a|} .$$

$\rho_{xy} = 1$ when $a > 0$

$\rho_{xy} = -1$ when $a < 0$.

(c) $$I(X,Y) = -E_p \ln \frac{p(X)\,p(Y)}{p(X,Y)} .$$

$I(X,Y) = 0$ implies that

$$p(X)\,p(Y) = p(X,Y)$$

We can then write

$$E(XY) = \iint xy\, p(x,y)\, dx\, dy$$

$$= \iint xy\, p(x)\, p(y)\, dx\, dy$$

$$= \int x\, p(x)\, dx \int y\, p(y)\, dy = E(X)\,E(Y)$$

$\therefore \quad E(XY) - E(X)E(Y) = 0$

$\Rightarrow cov(X, Y) = 0$

~~Hence $\rho(X,Y) = 0$~~

Hence, $\rho_{XY} = 0$

(d) NO, even if $\rho_{XY} = 0$, $I(X,Y)$ can be nonzero. Counterexample:
Let X, Y take values, $(1,0)$, $(0,1)$, $(-1,0)$, $(0,-1)$ with probability 0.25.

$E(X) = 0, \quad E(Y) = 0 \quad E(XY) = 0$

$\therefore \quad E(XY) - E(X).E(Y) = 0$

$cov(X, Y) = 0$

$$I(X,Y) = -\sum p(X,Y) \ln \frac{p(X)p(Y)}{p(X,Y)}$$

$$= -p(X=1, Y=0) \ln \frac{p(X=1)p(Y=0)}{p(X=1,Y=0)} - p(X=0,Y=1) \ln \frac{p(X=0)p(Y=1)}{p(X=0,Y=1)}$$

$$- p(X=-1, Y=0) \ln \frac{p(X=-1)p(Y=0)}{p(X=-1,Y=0)} - p(X=0, Y=-1) \ln \frac{p(X=0)p(Y=-1)}{p(X=0,Y=-1)}$$

$$= -\frac{1}{4} \ln \frac{\cancel{1/4} \cdot 1/2}{\cancel{1/4}} - \frac{1}{4} \ln \frac{1/2 \cdot \cancel{1/4}}{\cancel{1/4}} - \frac{1}{4} \ln \frac{\cancel{1/4} \cdot 1/2}{\cancel{1/4}}$$

$$-\frac{1}{4} \ln \frac{1/2 \cdot 1/4}{1/4}$$

$$= 4 \ln 2 \neq 0$$

Problem 2

30/30

2.1>

$$P(Y=1|X) = \frac{P(X|Y=1)\,P(Y=1)}{P(X|Y=1)\,P(Y=1) + P(X|Y=0)\,P(Y=0)}$$

$$= \frac{1}{1 + \frac{P(Y=0)}{P(Y=1)}\frac{P(X|Y=0)}{P(X|Y=1)}}$$

$$= \frac{1}{1 + \exp \ln\left[\frac{P(Y=0)}{P(Y=1)} \cdot \frac{P(X|Y=0)}{P(X|Y=1)}\right]}$$

$$= \frac{1}{1 + \exp\left[\ln \frac{(1-\pi)}{\pi} + \sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right]}$$

Now, we know that:

$$P(X_i|Y=k) = \frac{1}{\sigma_{ik}\sqrt{2\pi}} \exp \frac{-(X_i - \mu_{ik})^2}{2\sigma_{ik}^2}$$

$$\therefore \ln P(X_i|Y=k) = -\ln \sigma_{ik}\sqrt{2\pi} - \frac{(X_i - \mu_{ik})^2}{2\sigma_{ik}^2}$$

$$= -\ln \sigma_{ik}\sqrt{2\pi} - \frac{(X_i^2 - 2X_i\mu_{ik} + \mu_{ik}^2)}{2\sigma_{ik}^2}$$

Hence $k = 0, 1$,

We can write

$$\sum_i [\ln P(x_i \mid Y=0) - \ln P(x_i \mid Y=1)]$$

$$= \ln(\sigma_{i1}\sqrt{2\pi}) + \frac{x_i^2}{2\sigma_{i1}^2} - \frac{x_i \mu_{i1}}{\sigma_{i1}^2} + \frac{\mu_{i1}^2}{2\sigma_{i1}^2}$$

$$- \ln(\sigma_{i0}\sqrt{2\pi}) - \frac{x_i^2}{2\sigma_{i0}^2} + \frac{x_i \mu_{i0}}{\sigma_{i0}^2} + \frac{\mu_{i0}^2}{2\sigma_{i0}^2}$$

$$= \ln\frac{\sigma_{i1}}{\sigma_{i0}} + \frac{x_i^2}{2}\left(\frac{1}{\sigma_{i1}^2} - \frac{1}{\sigma_{i0}^2}\right) + x_i\left(\frac{\mu_{i0}^2}{\sigma_{i0}^2} - \frac{\mu_{i1}^2}{\sigma_{i1}^2}\right) + \left(\frac{\mu_{i1}^2}{2\sigma_{i1}^2} - \frac{\mu_{i0}^2}{2\sigma_{i0}^2}\right).$$

$$P(Y=1 \mid X) = \frac{1}{1 + \frac{(1-\pi)}{\pi}\exp\left[\sum_i \left\{ \ln\frac{\sigma_{i1}}{\sigma_{i0}} + \frac{x_i^2}{2}\left(\frac{1}{\sigma_{i1}^2} - \frac{1}{\sigma_{i0}^2}\right) + x_i\left(\frac{\mu_{i0}^2}{\sigma_{i0}^2} - \frac{\mu_{i1}^2}{\sigma_{i1}^2}\right) + \left(\frac{\mu_{i1}^2}{2\sigma_{i1}^2} - \frac{\mu_{i0}^2}{2\sigma_{i0}^2}\right)\right.\right.}$$

Therefore this model is NOT the form used by logistic regression. This is because the exponential term in the logistic function has an exponent <u>linear</u> in $x$, whereas, in this case, the exponent is a <u>quadratic</u> function of $x$.

2.2> Similar to previous case, let us calculate $P(Y=1|X)$

$$P(Y=1|X) = \frac{P(Y=1)\,P(X|Y=1)}{P(Y=1)\,P(X|Y=1) + P(Y=0)\,P(X|Y=0)}$$

$$= \frac{1}{1 + \frac{P(Y=0)\,P(X|Y=0)}{P(Y=1)\,P(X|Y=1)}}.$$

$$= \frac{1}{1 + \frac{(1-\pi)}{\pi} \exp \ln \frac{P(X|Y=0)}{P(X|Y=1)}}$$

now

$$\ln P(X|Y=k) = -\ln 2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}$$

$$-\left[\frac{\sigma_2^2(x_1-\mu_{1k})^2 + \sigma_1^2(x_2-\mu_{2k})^2 - 2\sigma_1\sigma_2(x_1-\mu_{1k})(x_2-\mu_{2k})}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\right]$$

$$= -\ln 2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}$$

$$-\left[\frac{x_1^2\sigma_2^2 + \sigma_1^2x_2^2 - 2\sigma_1\sigma_2x_1x_2}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\right]$$ does not depend on k

$$-\left[\frac{\sigma_2^2(-2x_1\mu_{1k} + \mu_{1k}^2) + \sigma_1^2(-2x_2\mu_{2k} - \mu_{2k}^2) - 2\sigma_1\sigma_2(\mu_{1k}\mu_{2k} - x_1\mu_{2k} - x_2\mu_{1k})}{2(1-\rho^2)\sigma_1^2\sigma_2^2}\right.$$ depends on k

The ~~first~~ first and second term does not depend on k. Hence when we write

$\ln P(X|Y=0) - \ln P(X|Y=1)$, the first and second terms cancel. We are left with:

$\ln P(X|Y=0) - \ln P(X|Y=1)$

$$= \frac{\sigma_2^2(-2X_1\mu_{11} + \mu_{11}^2) + \sigma_1^2(-2X_2\mu_{21} + \mu_{21}^2) - 2\sigma_1\sigma_2(\mu_{11}\mu_{21} - X_1\mu_{21} - X_2\mu_{11})}{2(1-\rho^2)\sigma_1^2\sigma_2^2}$$

$$- \frac{\sigma_2^2(-2X_1\mu_{10} + \mu_{10}^2) + \sigma_1^2(-2X_2\mu_{20} + \mu_{20})^2 - 2\sigma_1\sigma_2(\mu_{10}\sigma_{20} - X_1\mu_{20} - X_2\mu_{20})}{2(1-\rho^2)\sigma_1^2\sigma_2^2}$$

$$= \frac{X_1(-2\sigma_2^2\mu_{11} + 2\sigma_2^2\mu_{10} + 2\sigma_1\sigma_2(\mu_{21} - \mu_{20})) + X_2(-2\sigma_1^2\mu_{21} + 2\sigma_1^2\mu_{20}) + 2\sigma_1\sigma_2(\mu_{21} - \mu_{20}) - 2\sigma_1\sigma_2(\mu_{11}\mu_{21} - \mu_{10}\mu_{20}) + \sigma_2^2(\mu_{11}^2 - \mu_{10}^2) + \sigma_1^2(\mu_{21}^2 - \mu_{20}^2)}{2(1-\rho^2)\sigma_1^2\sigma_2^2}$$

$$= w_0' + w_1'X_1 + w_2'X_2$$

Therefore, we can finally write

$$P(Y=1|X) = \frac{1}{1 + \exp\left[\ln\left(\frac{1-\pi}{\pi}\right) + (w_0' + w_1' x_1 + w_2' x_2)\right]}$$

$$= \frac{1}{1 + \exp(w_0 + w_1 x_1 + w_2 x_2)}$$

This <u>IS</u> the form used by the logistic regression, since the exponent is linear in $(x_1, x_2)$.

PROBLEM 3

36/40

3.1> For each category, the no. of parameters we will need to estimate is

$$|V| \times |L| = 5 \times 10^4 \times 10^3$$
$$= 5 \times 10^7$$

Even if there are 10 categories, and each category has a 100 documents, we will only have $100 \times 1000 = 10^5$ words.
Hence, we have more $P(X_i|Y)$ to evaluate than the no. of words, and so, most of these estimates will be either 0, or inaccurate.

3.2> The overall testing accuracy is 78.521%
The confusion matrix is printed in the next page.

3.3> Adding up individual columns of the confusion matrix suggests some ~~cate~~ groups are more miscategorized than others.
"comp." groups are more miscategorized because I think they would contain similar content. (ibm vs mac vs windows)
Similarly electronics, since its a very

broad category, and has similarities to hardware groups in "comp."
"talk.politics.misc" and "talk.religion.misc" are also miscategorized a lot, since they contain broad miscellaneous topics.

3.4> The plot is attached in the next page. I took 10 points between $10^{-5}$ and 1 and the reported values are:

$10^{-5}$: 78.39%
$3.6\times10^{-5}$: 78.6%
$1.29\times10^{-4}$: 78.97%
$4.64\times10^{-4}$: 79.5%
$1.67\times10^{-3}$: 79.73%
$5.99\times10^{-3}$: 80.06%
$2.15\times10^{-2}$: 80.59%
$7.74\times10^{-2}$: 80.61%
$2.78\times10^{-1}$: 80.39%
$1.00$: 78.11%

At low values of $\alpha$, accuracy drops, because some $P(x_i/Y)$ are taken to be very ~~close~~ small simply because there are no training examples for those words and categories.

At high values of $\alpha$, the prior dominates over the evidence, hence likelihood estimates from ~~or~~ training examples are washed out by the prior, which is much bigger. So, accuracy drops.

3.5) The important metric here is $I(Y, X)$

For each word $X_j$ in the vacobaluary, we can define $I(Y, X_j)$ as

$$I(Y; X_j) = H(Y) - H(Y/X_j)$$

To calculate this metric, we ~~cal~~ can calculate

$$H(Y) = \sum_k P(Y = y_k) \ln P(y = y_k).$$

$$H(y/X_j) = P(X_j = 1)\, H(Y/X_j = 1) + P(X_j = 0)\, H(Y|X_j = 0).$$

First we calculate

$$P(X_j = 1) = \sum_k P(X_j = 1 / Y = y_k)\, P(Y = y_k)$$

$$P(X_j = 0) = 1 - P(X_j = 1)$$

Now, from Bayes rule,

$$P(Y = y_k | X_j = 1) = \frac{P(X_j = 1 / Y = y_k)\, P(Y = y_k)}{P(X_j = 1)}$$

We know the numerator from estimates made from traing examples. The denominator is calculated above.

Now, from these values we can calculate

$$H(Y|X_j = 1) = \sum_k P(Y = y_k | X_j = 1) \ln P(Y = y_k | X_j = 1).$$

Similarly, for $H(Y|X_j=0)$, we can write

$$P(X_j=0|Y=y_k) = 1 - P(X_j=1|Y=y_k)$$

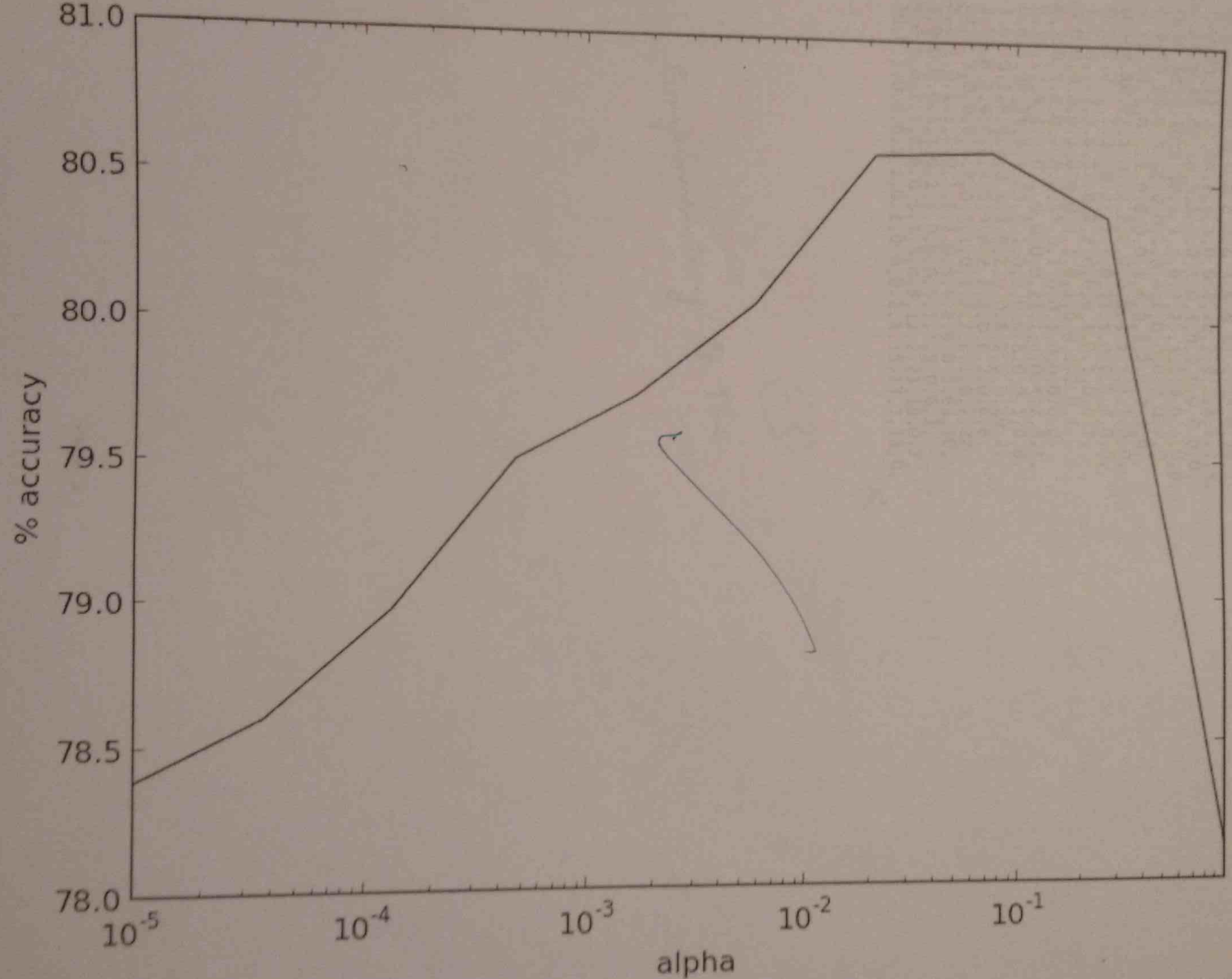
$$\therefore P(Y=y_k|X_j=0) = \frac{P(X_j=0|Y=y_k)\,P(Y=y_k)}{P(X_j=0)}$$

$$H(Y|X_j=0) = \sum_k P(Y=y_k|X_j=0) \ln P(Y=y_k|X_j=0)$$

3.6> The program calculates $I(Y, X_j)$ for each word $X_j$ in the vocabulary, and calculates the words with 100 highest $I(Y, X_j)$. These are printed in the next page.

3.7> 1) The dataset still contains often used words like 'of', 'is', 'we', etc. probably because longer documents containing lots of such words were chosen.

2) It also contains too many words related to computers, such as windows, monitor, ram. etc. This might introduce ~~introduce~~ some bias.


81.0
80.5
80.0
79.5
79.0
78.5
78.0
% accuracy
$10^{-5}$
$10^{-4}$
$10^{-3}$
$10^{-2}$
$10^{-1}$
alpha

```
0,0,1,0,0,0,0,0,0,0,2,0,3,5,0,11,1,12,6,39,
0,0,33,11,17,54,7,3,1,0,0,3,20,7,8,2,1,1,1,3,
0,13,0,30,13,16,5,1,0,0,1,0,4,0,0,0,0,0,0,0,
0,14,57,0,30,6,32,2,1,1,0,3,25,3,1,0,0,1,0,0,
0,9,19,20,0,3,16,0,0,1,0,4,7,0,0,0,0,0,1,0,
1,22,21,1,0,0,1,0,0,0,0,1,4,0,3,2,1,0,1,0,
0,4,4,10,12,1,0,14,2,2,2,0,8,3,1,1,1,1,0,0,
0,1,2,2,2,1,17,0,27,1,1,0,11,5,0,0,2,2,0,0,
1,1,3,1,2,3,8,17,0,2,2,0,6,4,1,0,1,0,0,1,
0,0,0,0,0,0,1,0,0,0,4,1,0,1,0,0,1,2,0,1,
0,1,0,1,0,0,2,0,0,17,0,1,0,0,1,0,0,0,0,0,
2,11,12,4,3,5,0,1,0,0,0,0,21,1,4,0,4,2,5,1,
0,8,5,32,21,3,7,13,3,1,0,2,0,8,6,0,0,1,0,0,
3,6,10,1,8,6,4,0,1,3,0,2,9,0,5,2,5,0,10,2,
3,10,8,2,4,4,6,4,0,3,0,2,7,8,0,0,2,0,6,6,
24,1,3,0,0,0,0,2,0,5,1,0,1,7,3,0,1,6,2,27,
2,2,1,0,1,1,2,0,1,2,2,9,3,6,2,0,0,3,63,10,
3,0,0,0,0,1,1,0,1,1,0,0,0,5,1,1,5,0,6,3,
4,0,5,0,1,1,2,6,0,5,1,5,0,8,12,2,23,18,0,7,
26,0,3,0,0,0,1,1,0,1,0,0,0,2,1,15,13,1,13,0,
```

-5

Complimentary to this

100words

windows god he scsi car drive space team dos bike file
of that mb game key mac jesus window dod hockey the
graphics card image his gun encryption sale apple
government season we games israel disk files ide controller
players shipping chip program was cars nasa win year were
they turkish motif people armenian play drives bible use
widget pc clipper offer jpeg baseball bus my nhl
software is db server jews os israeli output data
system who league armenians for christian christians
entry mhz ftp price christ guns thanks church color teams
privacy condition launch him com monitor ram

for metric $I(x_i; y)$.