# ML Homework 1 Solutions

Ahuva Bechhofer

February 17, 2023

## 1 Analyzing Bayes Classifier

### 1.1 i

To find the $P[Y = 1|A + B]$, we can use the exponential CDF since A and B are Known and C is the only unknown.
Since C's PDF was given to be exponential, the CDF will compute all possibilities for C that satisfy the constraints. Therefore $P[Y = 1|A+B] = 1-e^{-7+a+b}$.
To find the Bays we follow the definition of a bays classifier
$argmx_{y\epsilon 0,1} P[Y = y|X = \vec{x}]$
this can be further expanded and simplified to

$$P[Y = 1|x] \geq P[Y = 0|x] \rightarrow [Y = 1|x] \geq 0.5 \tag{1}$$

Since we know

$$[Y = 1|x] = 1 - e^{-7+a+b} \tag{2}$$

we can set

$$1 - e^{-7+a+b} \geq 0.5 \tag{3}$$

and solve for a+b.
Doing that computation we get

$$a + b \leq 7 + ln(0.5). \tag{4}$$

Therefore our bays classifier is as follows:

$$f^*(a, b) = \begin{cases} 1 & \text{if } a + b \leq 7 + ln(.5) \\ 0 & \text{if } a + b \geq 7 + ln(.5) \end{cases} \tag{5}$$

To find the bays error we compute $P[Y \neq y]$
let z = A+B by the law of total probability we want to compute

$$\int_z P_z(z) * P[f^*(x) = y|Z = z]$$

$P[f^*(x) = y|Z = z]$ can be broken down to $P[Y = 1, f^*(x) = 0|Z = z]$ and to $P[Y = 0, f^*(x) = 1|Z = z]$
In addition, $P(Z)$ is the sum of two exponentials, so in order to compute that we have to figure out what the distribution for that looks like.
If we set for example two points $A+B = \alpha$ we can find $P[B = \alpha - A|A = a] * P[A]$ but this would be only for a particular point, to do this for all points, we have

$$P(\alpha) = \int_0^\alpha P(\beta) * P[\alpha - \beta]\, d\beta$$

Since we know A and B come from the exponential distribution with $\lambda = 1$ we have

$$\int_0^\alpha e^{-\beta} * e^{-(\alpha-\beta)}\, d\beta$$

when solved we get $e^{-\alpha} * \alpha$ as the distribution of two exponentials.
Now we can go back to the error approximation, if we choose the case $f^*(x) = 1$ we get

$$\int_0^{7+ln(0.5)} P_z(z) * P[Y = 0|Z = z]$$

and for the case $f^*(x) = 0$

$$\int_{7+ln(0.5)}^{7} P_z(z) * P[Y = 1|Z = z]$$

Adding the two cases and substituting the distribution in we get

$$\int_0^{7+ln(0.5)} ze^{-z} * e^{-7+a+b}\, dz + \int_{7+ln(0.5)}^{7} ze^{-z} * (1 - e^{-7+a+b})\, dz \qquad (6)$$

which when evaluated this equals 0.0199612

## 1.2

Following the same logic of part A we have to find the distribution of two exponentials since here only A is known. But since we already computed that in part a, we can now integrate the PDF we computed to find $P[Y = 1|A]$
If we set $B + C = \alpha$ following our constraints we get $\alpha \leq 7 - A$, therefore when integrating that is our upper bound. So we set

$$\int_0^{7-A} \alpha * e^{-\alpha}\, d\alpha$$

solving this we get $e^{-7+A}[A-8]+1$ To find the bays we follow exactly what was done in the previous part so we have $e^{-7+A}[A-8]+1 \geq 0.5$ using Wolfram Alpha

to solve for A, we get $A \leq 5.32165$ therefore $f^*(A) = \begin{cases} 1 & \text{if } A \leq 5.32165 \\ 0 & \text{if } 5.32165 \leq A \leq 7 \end{cases}$
To find the error again we have to find $P[Y \neq y|A]$ we can again break it down to the same cases as the previous part and we end up with

$$\int_{5.32165}^{7} P[A] * P[Y = 0|A] + \int_{0}^{5.32165} P[A] * P[Y = 1|A] \tag{7}$$

since here A is known to be distributed exponentially we can just use the exponential PDF for P[A]' So we have

$$\int_{5.32165}^{7} e^{-A} * (1 - (e^{-7+A}[A - 8] + 1))\, dA + \int_{0}^{5.32165} e^{-A} * (e^{-7+A}[A - 8] + 1)\, dA \tag{8}$$

which equals 0.0270683

## 1.3

Since here none of A,B or C are known we have to find the distribution of 3 exponentials, but since we already have the distribution for two exponentials we do the same operation (known as convolution) on the sum of two exponential plus another exponential.
so we get

$$P_z(z) = \int_0^z u * e^{-u} * e^{-(z-u)}\, du = \frac{e^{-\alpha} * \alpha^2}{2}$$

Now to find the bays we just integrate the this expression from 0 to 7 and we get 0.970364. So the bays will always output 1 since this probability is greater than 0.5. In addition the error $P[Y \neq y]$ is 1 - 0.970364

## 1.4 ii

Since A and B are known but C is not we want to pick a distribution for C that "overrides" A and B, therefore if we pick the distribution of C to be $\Delta$ or + $\Delta$ where $\Delta$ is a very large number, the value of A+B won't matter since it would be insignificant in terms of C and A and B won't affect the $A + B + C \leq 7$ inequality.
To calculate the error we take $\lim_{\Delta\to\infty} P[Y \neq y]$ if we break it into the two cases $\lim_{\Delta\to\infty} 1[Y = 1] + 1[Y = 0]$ which shows that $\lim_{\Delta\to\infty} P[Y = 0] = 0.5$ and $\lim_{\Delta\to\infty} P[Y = 1] = 0.5$ which means P[error] = 0.5. This is because as we described at the beginning as $\Delta$ increases, the larger it gets the more it "overrides" A and B and since our distribution for $\Delta$ only has two options and one of the cases will be chosen we get a 50 percent probability of being chosen and 50 percent of not being chosen i.e. the error.

# 2 Classification with Asymmetric Costs

## 2.1

In the real world some mistakes are more expensive than others, therefore lets say for a model that does classifies diseases some diseases are way worse than others, so those should have a high cost to indicate the importance and penalize if an error is made. In addition having the model output a cost of a lack of confidence is also useful since it can help with the evaluation of the confidence of the model and humans can intervene if a model says it is not sure about a diagnosis, so in the above example so a doctor can interject to help with a diagnosis.

## 2.2

To start we compute the expectation of the loss function.

$$E[l(x,y)|f(x) = 0, 1, -1, X = \vec{x}] \tag{9}$$

lets break this into cases:
f(x) = 0:

$$p * P[f(x) = 0, Y = 1|X = \vec{x}] \rightarrow p[1[f(x) = 0]P[Y = 1|X = \vec{x}]] \tag{10}$$

since this is the f(x) = 0 case the indicator function returns one for that and this can be simplified to p $\eta(\vec{x})$

for f(x) = 1:

$$q * P[f(x) = 1, Y = 0|X = \vec{x}] \rightarrow q[1[f(x) = 1]P[Y = 0|X = \vec{x}]] \tag{11}$$

Since this is the f(x) = 1 case the indicator function returns one for that and this can be simplified to $q(1 - \eta(\vec{x}))$

for $f(x) = -1 : r * P[f(x) = -1|X = \vec{x}] \rightarrow r * 1[f(x) = -1]](12)$

since this is the $f(x) = -1$ case the indicator function returns one for that and this can be simplified to r

Since we are looking to prove $f^*(x)$ is the bays, we know it will pick the option that will minimize the loss, so if we are at case one, and the bays picks zero then we know that that MUST be less than the loss of the other cases. therefore we can set up two inequalities for each case and solve for the bounds.

case of 0:

$$p[\eta(\vec{x})] \leq q[1 - \eta(\vec{x})], p[\eta(\vec{x})] \leq r \tag{13}$$

We used algebra to isolate $\eta(\vec{x})$ and then we got by using the algebra and the given assumption $r < \frac{pq}{p+q}$ that the bound is $0 \leq \eta(\vec{x}) \leq \frac{r}{q}$
0 is a lower bound for $\eta(\vec{x})$ since we cannot have a negative probability.

case of -1:

$$r \leq \eta(\vec{x}), r \leq q(1 - \eta(\vec{x}) \tag{14}$$

again by using algebra we isolate $\eta(\vec{x})$ and we get $\frac{r}{p} \leq \eta(\vec{x}) \leq 1 - \frac{r}{p}$

case of 1:

$$q[1 - \eta(\vec{x})] \leq p[\eta(\vec{x})], q[1 - \eta(\vec{x})] \leq r \tag{15}$$

by doing the same as above we get the bounds $\eta(\vec{x}) \geq 1 - \frac{r}{q}$ and the upper limit of $\eta(\vec{x})$ is 1 since this is a probability and thus we cannot have a probability greater than one.

## 2.3

we do exactly as above.
case of 0:

$$P[\eta(\vec{x})] \leq q[1 - \eta(\vec{x})] \tag{16}$$

We used algebra to isolate $\eta(\vec{x})$ and then we got by using the algebra and the given assumption $r > \frac{pq}{p+q}$ that the bound is $0 \leq \eta(\vec{x}) \leq \frac{q}{p+q}$
0 is a lower bound since we cannot have a negative probability.

case of 1:

$$q[1 - \eta(\vec{x})] \leq r, q[1 - \eta(\vec{x})] \leq p * \eta(\vec{x}) \tag{17}$$

with algebra we isolate eta and get the bound $\frac{q}{p+q} \leq \eta(\vec{x})] \leq 1$ as the same reasons as the previous part.

## 2.4

if set p=q for the bounds in the last part we get $f^*(x) = \begin{cases} 0 & \text{if } 0 \leq \eta(\vec{x}) \leq \frac{p}{2p} \\ 1 & \text{if } \frac{p}{2p} \leq \eta(\vec{x}) \leq 1 \end{cases}$ which then becomes $f^*(x) = \begin{cases} 0 & \text{if } 0 \leq \eta(\vec{x}) \leq \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq \eta(\vec{x}) \leq 1 \end{cases}$ same idea is applied in the part beforehand, we plug into $r\frac{p}{2}$ and that simplifies when we have $\frac{r}{p}to\frac{1}{2}$ so the bounds for that are again $f^*(x) = \begin{cases} 0 & \text{if } 0 \leq \eta(\vec{x}) \leq \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq \eta(\vec{x}) \leq 1 \\ -1 & \text{if } \frac{1}{2} \leq \eta(\vec{x}) \leq \frac{1}{2} \end{cases}$ We can get rid of the

-1 case since there is no number between itself. Now what we have is exactly the definition of a bayes classifier. this makes sense since now were removing the asymmetric cost so by definition of bayes we would get the min loss.

# 3 Finding (local) minima of generic functions

## 3.1

we are given $|f'(a) - f'(b)| \leq L^2|a - b|$ To prove the bound of the second derivative, we can divide by $|a - b|$ and take the limit as such:

$\lim_{a\to b}\left|\frac{f'(a)-f'(b)}{a-b}\right| \leq L^2$ (18)

this is the exact definition of a derivative (it can be seen more clearly if you set $z = a - b$ and put it in therms of that) but none the less now we get $f''(z) \leq L^2$

using the taylor remainder theorem we set $a = x$ and $b - a = \eta f'(x)$ and sub those values into the remainder equation.
$f(\bar{x}) = f(x) + f'(x)(-\eta f'(x)) + \frac{1}{2} * f''(z)(-\eta f'(x))^2$ now we subtract $f(x)$ to move it to the other side and combine terms so we get
$f(\bar{x}) - f(x) = (-\eta f'(x)^2) + \frac{1}{2} * f''(z)(\eta f'(x))^2$
since we know $L^2 \geq f''(z)$ we can write

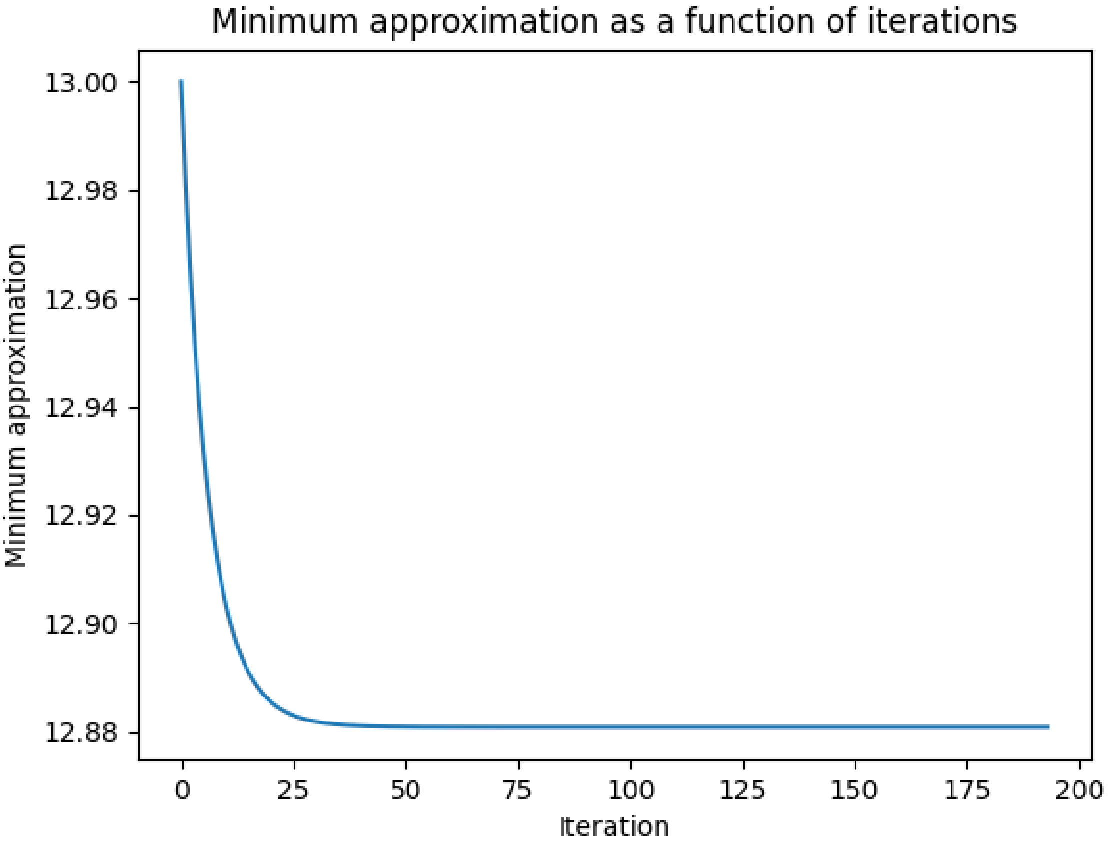
$$f(\bar{x})-f(x) \leq (-\eta f'(x)^2)+\frac{1}{2}*L^2(\eta f'(x))^2 \rightarrow f(\bar{x})-f(x) \leq (\eta f'(x)^2)(\frac{\eta}{2}*L^2-1) \tag{19}$$

Since we assumed $L \geq 0$ and were stating there exists some $\eta > 0$ we see that the $(\eta f'(x)^2)$ term will always be positive and that $\frac{\eta}{2} * L^2$ will also always be positive. So in order for $f(\bar{x}) - f(x) < 0$ that mean $\frac{\eta}{2} * L^2 < 1 \rightarrow \eta < \frac{2}{L^2}$
This proved the $<$ part, for the '=' part since we cannot divide by zero the only case to get $f(\bar{x}) = f(x)$ is when $f'(x) = 0$

## 3.2

```
$tolerance = e^{-20}$
while($f(\bar{x_i}) - f(\bar{x_{i-1}}) \geq tolerance$) do:
$\bar{x} = x - \eta f'(x)$
return $\bar{x}$
```

### 3.3



### 3.4

Since gradient decent converges to the first minima if that happens to not be the global minima, it cannot change directions and go up and look for another minima. Therefore an idea I learnt in AI last semester is to start at random points and then maybe one of those points will lead the gradient to the global minima.

# 4 Exploring the limits Current Language Models

## 4.1 i

We will use the probabilities given to us for the bigram to calculate $P(Y|w_{1:n})$

$$\begin{aligned} P(Y|w_{1:n}) &= \frac{P(w_{1:n}|y)P(y)}{P(w_{1:n})} \\ &= \frac{P(w_{1:n}|y)P(y)}{\sum_y P(y)P(w_{1:n}|y)} \end{aligned}$$

since y represents the class either human or Chat GPT we will have two probabilities

$$P(human|w_{1:n}) = \frac{\prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2})P(human)}{\sum_{human} P(human)P(w_i|w_{i-1}w_{i-2}, human) + \sum_{GPT} P(GPT)P(w_i|w_{i-1}w_{i-2}, GPT)}$$

$$P(GPT|w_{1:n}) = \frac{\prod_{i=1}^{n} P(w_i|w_{i-1}w_{i-2})P(GPT)}{\sum_{human} P(human)P(w_i|w_{i-1}w_{i-2}, human) + \sum_{GPT} P(GPT)P(w_i|w_{i-1}w_{i-2}, GPT)}$$

Now just sub $P(w_i|w_{i-1}w_{i-2}) = \frac{C(w_{i-1}w_{i-2}w_i)+1}{C(w_{i-1}w_{i-2})+|V|}$

$$P(human|w_{1:n}) = \frac{\prod_{i=1}^{n} \frac{C(w_{i-1}w_{i-2}w_i)+1}{C(w_{i-1}w_{i-2})+|V|})P(human)}{\sum_{human} P(human)\frac{C(w_{i-1}w_{i-2}w_i)+1}{C(w_{i-1}w_{i-2})+|V|}) + \sum_{GPT} P(GPT)\frac{C(w_{i-1}w_{i-2}w_i)+1}{C(w_{i-1}w_{i-2})+|V|}}$$

$$P(GPT|w_{1:n}) = \frac{\prod_{i=1}^{n} \frac{C(w_{i-1}w_{i-2}w_i)+1}{C(w_{i-1}w_{i-2})+|V|}P(GPT)}{\sum_{human} P(human)\frac{C(w_{i-1}w_{i-2}w_i)+1}{C(w_{i-1}w_{i-2})+|V|} + \sum_{GPT} P(GPT)\frac{C(w_{i-1}w_{i-2}w_i)+1}{C(w_{i-1}w_{i-2})+|V|}}$$

## 4.2 b

For the OOV rate I got around 9 percent for the bigram and 33 percent for the trigram.

## 4.3 c

My code is outputting funcky numbers for the probabilities and I don't think I have time to debug why. I know on a high level the Bigram should preform better since it does have a lower OOV rate and thus it can compute more of the word probabilities.

Note I did read an article online which connected the probabilities of N-gram and code execution since I was confused where to start when I first started the coding assignment. "https://medium.com/codex/statistical-language-model-n-gram-to-calculate-the-probability-of-word-sequence-using-python-2e54a1084250"