ML Homework 1

Maitar Asher

2023-02-16

Problem 1 - Analyzing Bayes Classifier

(i) (a)

$$P(Y=1|A=a,B=b) = P(c < (7-(a+b))$$

We can look at the CDF of an exponential random variable to calculate the probability that the random variable is less than $(7-(a+b))$

$$P(Y=1|A=a,B=b) = F((7-(a+b),1) = 1-e^{-1(7-(a+b))} = 1-e^{-7+(a+b)}$$

To construct the classifier we will need to check when
$P(Y=1|A=a,B=b) > 0.5$

$$1-e^{-7+(a+b)} > 0.5$$

$$e^{-7+(a+b)} < 0.5$$

$$e^{-7+(a+b)} < e^{ln(0.5)}(e^{ln(x)} = x)$$

$$-7+(a+b) < ln(0.5)$$

$$(a+b) < ln(0.5)+7$$

$$\hat{y}(a+b) = \begin{cases} 1, & \text{if } (a+b) < ln(0.5)+7 \\ 0, & \text{otherwise} \end{cases}$$

To find the Bayes error we would need to find what is the probability of our classifier predicting the wrong class.
Let z be denoting $a+b$, we would need to look at all possible values of z between 0 and 7 because when z is greater than 7 our classifier will definitely return 0

$$P(\hat{y}(z) \neq y) = \int_0^7 P(z) \cdot \sum_{y \in Y} (\mathbb{1}[\hat{y}(z) = y] \cdot P(Y \neq y|z))\, dz$$

$$= \int_0^7 P(z) \cdot (\mathbb{1}[\hat{y}(z) = 0] \cdot P(Y=1|z) + \mathbb{1}[\hat{y}(z) = 1] \cdot P(Y=0|z))\, dz$$

$$= \int_0^{ln(0.5)+7} P(z) \cdot (\mathbb{1}[\hat{y}(z) = 0] \cdot P(Y = 1|z) + \mathbb{1}[\hat{y}(z) = 1] \cdot P(Y = 0|z))\, dz$$

$$+ \int_{ln(0.5)+7}^{7} P(z) \cdot (\mathbb{1}[\hat{y}(z) = 0] \cdot P(Y = 1|z) + \mathbb{1}[\hat{y}(z) = 1] \cdot P(Y = 0|z))\, dz$$

When z is between 0 and ln(0.5)+7 we know our classifier will always return 1, and when z is between ln(0.5)+7 and 7 we know our classifier will always return 0.

$$P(\hat{y}(z) \neq y) = \int_0^{ln(0.5)+7} P(z) \cdot P(Y = 0|z)\, dz + \int_{ln(0.5)+7}^{7} P(z) \cdot P(Y = 1|z)\, dz$$

To calculate P(z) we would need to check what is the probability of the summation of two i.i.d exponential random variables taking a specific value. credit

$$f(X_1 + X_2, \lambda) = \lambda^2 (X_1 + X_2) e^{-\lambda(X_1+X_2)}$$

Then $P(z) = ze^{-z}$. Plugging all in we get:

$$P(\hat{y}(z) \neq y) = \int_0^{ln(0.5)+7} ze^{-z} \cdot (e^{-7+z})\, dz + \int_{ln(0.5)+7}^{7} ze^{-z} \cdot (1 - e^{-7+z})\, dz$$

$$= \int_0^{ln(0.5)+7} e^{-7} z\, dz + \int_{ln(0.5)+7}^{7} ze^{-z} - e^{-7} z\, dz$$

Using integration by parts:

$$\int ze^{-z}\, dz = -ze^{-z} - \int (-e^{-z})\, dz = -ze^{-z} - e^{-z}$$

Then:

$$P(\hat{y}(z) \neq y) = \left(e^{-7} \frac{z^2}{2} \right) \Bigg|_0^{ln(0.5)+7} + \left(-ze^{-z} - e^{-z} - e^{-7} \frac{z^2}{2} \right) \Bigg|_{ln(0.5)+7}^{7}$$

$$= \left(e^{-7} \cdot \frac{(ln(0.5) + 7)^2}{2} \right) - (0)$$

$$+ \left(-7e^{-7} - e^{-7} - e^{-7} \cdot \frac{7^2}{2} \right) - \left(-(ln(0.5) + 7)e^{-(ln(0.5)+7)} - e^{-(ln(0.5)+7)} - e^{-7} \cdot \frac{(ln(0.5) + 7)^2}{2} \right)$$

$$= -32.5 \cdot e^{-7} + (ln(0.5) + 8) \cdot e^{-(ln(0.5)+7)} + e^{-7} \cdot (ln(0.5) + 7)^2 \approx 0.0199$$

(i) (b)

$$P(Y = 1|A = a) = P((b + c) < (7 - a))$$

We can look at the CDF of the summation of 2 i.i.d exponential random variables to calculate the probability that the sum is less than $(7 - a)$
credit

$$F(X_1 + X_2, \lambda) = 1 - e^{-\lambda(X_1+X_2)} - \lambda(X_1 + X_2)e^{-\lambda(X_1+X_2)}$$

$$P(Y = 1|A = a) = F((7-a), 1) = 1-e^{-(7-a)}-(7-a)\cdot e^{-(7-a)} = 1-(8-a)\cdot e^{-7+a}$$

To construct the classifier we will need to check when
$P(Y = 1|A = a) > 0.5$

$$1 - (8 - a) \cdot e^{-7+a} > 0.5$$

$$(8 - a) \cdot e^{-7+a} < 0.5$$

$$a > 7.76804, a < 5.32165$$

$a > 7.76804$ is canceled because we know our classifier should return 0 when $(a + b + c \geq 7)$

$$\hat{y}(a) = \begin{cases} 1, & \text{if } a < 5.32165 \\ 0, & \text{otherwise} \end{cases}$$

To find the Bayes error we would need to find what is the probability of our classifier predicting the wrong class.
We would need to look at all possible values of a between 0 and 7 because if a is greater than 7 our classifier will definitely return 0

$$P(\hat{y}(a) \neq y) = \int_0^7 P(a) \cdot \sum_{y \in Y} (\mathbb{1}[\hat{y}(z) = y] \cdot P(Y \neq y|a))\, da$$

$$= \int_0^7 P(a) \cdot (\mathbb{1}[\hat{y}(a) = 0] \cdot P(Y = 1|a) + \mathbb{1}[\hat{y}(a) = 1] \cdot P(Y = 0|a))\, da$$

$$= \int_0^{5.32165} P(a) \cdot (\mathbb{1}[\hat{y}(a) = 0] \cdot P(Y = 1|a) + \mathbb{1}[\hat{y}(a) = 1] \cdot P(Y = 0|a))\, dz$$

$$+ \int_{5.32165}^7 P(a) \cdot (\mathbb{1}[\hat{y}(a) = 0] \cdot P(Y = 1|a) + \mathbb{1}[\hat{y}(a) = 1] \cdot P(Y = 0|a))\, da$$

When a is between 0 and 5.32165 we know our classifier will always return 1, and when a is between 5.32165 and 7 we know our classifier will always return 0. Then,

$$P(\hat{y}(a) \neq y) = \int_0^{5.32165} P(a) \cdot P(Y = 0|a)\, da + \int_{5.32165}^7 P(a) \cdot P(Y = 1|a)\, da$$

To calculate P(a) we need to check what is the probability of one exponential random variable taking a specific value. Using PDF of exponential random variable we get: $P(a) = e^{-a}$. Plugging all in we get:

$$P(\hat{y}(a) \neq y) = \int_0^{5.32165} e^{-a}\cdot((8-a)\cdot e^{-7+a})\,da + \int_{5.32165}^{7} e^{-a}\cdot(1-(8-a)\cdot e^{-7+a})\,da$$

$$= \int_0^{5.32165} (8-a)\cdot e^{-7})\,da + \int_{5.32165}^{7} e^{-a} - (8-a)\cdot e^{-7})\,da$$

$$= \Big(e^{-7}\cdot(8a - \frac{a^2}{2})\Big)\Big|_0^{5.32165} + \Big(-e^{-a} - e^{-7}\cdot(8a - \frac{a^2}{2})\Big)\Big|_{5.32165}^{7}$$

$$\Big(e^{-7}\cdot(8\cdot 5.32165 - \frac{5.32165^2}{2})\Big) - (0)$$

$$+\Big(-e^{-7} - e^{-7}\cdot(8\cdot 7 - \frac{7^2}{2}\Big) - \Big(-e^{-5.32165} - e^{-7}\cdot(8\cdot 5.32165 - \frac{5.32165^2}{2}\Big) \approx 0.027$$

(i) (c)

$$P(Y=1) = P(a+b+c<7)$$

The summation of 3 i.i.d exponential random variables is equal to the distribution of gamma(n,λ) where λ is the min value of all exponential random variables and n is the number of variables we're summing.
Thus $A+B+C\ \lambda(3,1)$
Let x denote a + b + c, the PDF of x is:

$$f(x) = \frac{1^3\cdot x^{(3-1)}\cdot e^{-x}}{2!} = \frac{x^2\cdot e^{-x}}{2}$$

To find the CDF we can take the integral of the PDF between x to 0. Using integration by parts we get:

$$F(x) = \int_0^x \frac{x^2\cdot e^{-x}}{2}\,dx = \left(\frac{x^2}{2}\cdot(-e^{-x})\right)\Big|_0^x - \int_0^x(-xe^{-x})\,dx$$

$$= \left(\frac{x^2}{2}\cdot(-e^{-x})\right)\Big|_0^x - \left((x\cdot e^{-x})\,|_0^x - \int_0^x e^{-x}\,dx\right)$$

$$\left(\frac{x^2}{2}\cdot(-e^{-x})\right)\Big|_0^x - ((x\cdot e^{-x}) + e^{-x})\,|_0^x$$

$$= \left(-e^{-x}\left(\frac{x^2}{2}+x+1\right)\right)\Big|_0^x$$

$$-e^{-x}\left(\frac{x^2}{2}+x+1\right) + e^0 = -e^{-x}\left(\frac{x^2}{2}+x+1\right)+1$$

Then,

$$P(Y=1) = F(7) = -e^{-7}\left(\frac{7^2}{2}+7+1\right)+1 \approx 0.97$$

There is no randomness when A,B,C are unknown. Our classifier will return the class of the highest probability.

$$\hat{y} = \arg\max_{y \in Y} P(Y = y) = 1$$

There is no randomness in the classifier, so the Bayes error would be when A,B,C are not smaller than 7, thus

$$P(\hat{y} \neq 1) = 1 - (\approx)0.97 \approx 0.03$$

(ii)
Let C be a random variable that can take only two values, either some positive large value x or $-x$, each occurring with probability 0.5
Because A and B are known we are interested in computing the probability that $\mathbb{P}(C < 7 - (A + B)) = P(Y = 1)$. Because c can take only two values, the probability of c being smaller or greater than this expression will always be 0.5. If x approaches infinity, it does not matter what the value of (A+B) is, and essentially C overwrites (A+B). Thus we can see the limit of the Bayes error can be made as close to 0.5 as desired.

$$P(\hat{y}(a+b) \neq y) = \int_{-\infty}^{\infty} P(a+b) \cdot \sum_{y \in Y} (\mathbb{1}[\hat{y}(a+b) = y] \cdot P(Y \neq y | a+b)) \, d(a+b)$$

$$= \int_{-\infty}^{\infty} P(a+b) \cdot (\mathbb{1}[\hat{y}(a+b) = 0] \cdot P(Y = 1 | a+b) + \mathbb{1}[\hat{y}(a+b) = 1] \cdot P(Y = 0 | a+b)) \, d(a+b)$$

$$\lim_{x \to \infty} \mathbb{P}(\hat{y}(a+b) \neq y) = 0.5$$

Problem 2 - Classification with Asymmetric Costs

(i) This model of classification is more suitable for task problems that seek to provide different weighting for specific misclassification errors– i.e false positive/ false negative or overall lack of confidence. E.g a cancer diagnosis classifier that misclassifies a patient as healthy when they actually have a disease (false negative) or misclassifies a healthy person as having a disease (false positive) should not treat these errors uniformly in order to provide a more fitted error metric. Likewise, it should take into account the situations where the classifier does not know what to predict.

General analysis for (ii) and (iii)
Given the definition of the loss function, we get:

$$l(g(x), y) = p \cdot \mathbb{1}[g(x) = 0] \cdot \mathbb{1}[y = 1] + q \cdot \mathbb{1}[g(x) = 1] \cdot \mathbb{1}[y = 0] + 1 \cdot \mathbb{1}[g(x) = -1]$$

$$\mathbb{E}_{x,y}[l(g(x), y)] = \sum_{g(x),y} l(g(x), y) \cdot \mathbb{P}(g(x), y)$$

Let's fix x to simplify the calculation and our analysis. We can then integrate over x to remove the condition

$$\mathbb{E}_{y|x}[l(g(x), y)|x] = \sum_{g(x),y} l(g(x), y) \cdot \mathbb{P}(g(x), y|x)$$

$$= p \cdot \mathbb{P}(g(x) = 0, Y = 1|x) + q \cdot \mathbb{P}(g(x) = 1, Y = 0|x) + r \cdot \mathbb{P}(g(x) = -1|x)$$

There is no randomness in the probability of g(x) given x, then:

$$\mathbb{E}_{y|x}[l(g(x), y)|x] = p \cdot \mathbb{1}[g(x) = 0] \cdot \mathbb{P}(Y = 1|x) + q \cdot \mathbb{1}[g(x) = 1] \cdot \mathbb{P}(Y = 0|x) + r \cdot \mathbb{1}(g(x) = -1)$$

Given $n(x) = \mathbb{P}[Y = 1|X = x]$, we can plug it in:

$$\mathbb{E}_{x|y}[l(g(x), y)|x] = p \cdot \mathbb{1}[g(x) = 0] \cdot n(x) + q \cdot \mathbb{1}[g(x) = 1] \cdot (1 - n(x)) + r \cdot \mathbb{1}(g(x) = -1)$$

Now to show that the expected loss of $f^*(x)$ is less than $g(x)$ let us examine what is the difference between these expected values.

$$\mathbb{E}_{y|x}[l(f^*(x), y)|x] - \mathbb{E}_{x|y}[l(g(x), y)|x]$$

$$= p \cdot n(x) \cdot (\mathbb{1}[f^*(x) = 0] - \mathbb{1}[g(x) = 0]) + q \cdot (1 - n(x)) \cdot (\mathbb{1}[f^*(x) = 1] - \mathbb{1}[g(x) = 1])$$

$$+ r \cdot (\mathbb{1}[f^*(x) = -1] - \mathbb{1}[g(x) = -1])$$

For $f^*(x)$ to perform better, the above expression should be smaller or equal to 0 (the loss of $g(x)$ is greater). For the cases in which both classifiers agree, we can conclude there is no difference. We should check however for the cases where the classifiers disagree and attempt to show that even when maximizing the linear expression it is still smaller or equal to 0. Our cases can be derived from the possible values of $n(x)$.

(ii)
Given $r < \frac{pq}{p+q}$ and $p, q, r > 0$

- $0 \leq n(x) \leq \frac{r}{p}$

    1. $f^*(x) = 0$ and $g(x) = 1$

       $p \cdot \frac{r}{p}(1) + q(1 - \frac{r}{p})(-1) - r(0) = r - q + \frac{rq}{p} < \frac{pq}{p+q} - q + \frac{\frac{pq^2}{p+q}}{p} = \frac{pq}{p+q} - q + \frac{q^2}{p+q} = \frac{pq-qp-q^2+q^2}{p+q} = 0 \leq 0$

    2. $f^*(x) = 0$ and $g(x) = -1$

       $p \cdot \frac{r}{p}(1) + q(1 - \frac{r}{p})(0) + r(-1) = r - r = 0 \leq 0$

- $\frac{r}{p} < n(x) < 1 - \frac{r}{q}$

    1. $f^*(x) = -1$ and $g(x) = 0$

       $p(1 - \frac{r}{q})(-1) + q(0) + r(1) = -p + \frac{rp}{q} + r < -p + \frac{p^2q}{(p+q)q} + \frac{pq}{p+q} = \frac{-p^2-pq+p^2+pq}{p+q} = 0 \leq 0$

    2. $f^*(x) = -1$ and $g(x) = 1$

       $p(0) + q(-1)(1 - (1 - \frac{r}{q})) + r(1) = -r + r = 0 \leq 0$

- $1 - \frac{r}{q} \leq n(x) \leq 1$

    1. $f^*(x) = 1$ and $g(x) = 0$

       $p(-1)(1-\frac{r}{q})+q(1)(1-(1-\frac{r}{q}))+r(0) = -p+\frac{pr}{q}+r < -p+\frac{\frac{p^2q}{p+q}}{q}+\frac{pq}{p+q} = -p + \frac{p^2}{p+q} + \frac{pq}{p+q} = \frac{-p^2-pq+p^2+pq}{p+q} = 0 \leq 0$

    2. $f^*(x) = 1$ and $g(x) = -1$

       $p(0) + q(1)(1 - (1 - \frac{r}{q})) + r(-1) = r - r = 0 \leq 0$

(iii)
Given $r \geq \frac{pq}{p+q}$ and $p, q, r > 0$

- $0 \leq n(x) < \frac{q}{p+q}$

    1. $f^*(x) = 0$ and $g(x) = 1$

       $p(\frac{q}{p+q})(1)+q(1-\frac{q}{p+q})(-1)+r(0) = \frac{pq}{p+q} - q + \frac{q^2}{p+q} = \frac{pq-qp-q^2+q^2}{p+q} = 0$
       thus < 0

2. $f^*(x) = 0$ and $g(x) = -1$

   $p(\frac{q}{p+q})(1) + q(0) + r(-1) = (\frac{pq}{p+q}) - r = (\frac{pq}{p+q}) - (\frac{pq}{p+q}) = 0$
   r is greater or equal to $\frac{pq}{p+q}$ thus total expression < 0

- $\frac{q}{p+q} \leq n(x) \leq 1$

  1. $f^*(x) = 1$ and $g(x) = 0$

     $p(\frac{q}{p+q})(-1)+q(1-\frac{q}{p+q})(1)+r(0) = -\frac{pq}{p+q}+q-\frac{q^2}{p+q} = \frac{-pq+qp+q^2-q^2}{p+q} = 0$
     thus ≤ 0

  2. $f^*(x) = 1$ and $g(x) = -1$

     $p(0) + q(1 - \frac{q}{p+q})(1) + r(-1) = q - \frac{q^2}{p+q} - r = \frac{qp+q^2-q^2}{p+q} - r = (\frac{pq}{p+q}) - (\frac{pq}{p+q}) = 0$
     r is greater or equal to $\frac{pq}{p+q}$ thus total expression ≤ 0

(iv) Now lets analyze our model when $p = q$ and $r > \frac{p}{2}$
For the classifier in part (ii) we can discern that it will never return -1, and that since $r > \frac{p}{2}$, it will return 1 for n(x) higher or equal to 0.5 ($\frac{r}{p} > 0.5$) :

$$f^*(x) = \begin{cases} 0, & \text{if } 0 \leq n(x) < \frac{r}{p} \\ -1, & \text{if } \frac{r}{p} < n(x) < 1 - \frac{r}{p} \\ 1, & \text{if } 1 - \frac{r}{p} \leq n(x) \leq 1 \end{cases}$$

For the classifier in part (iii) we get:

$$f^*(x) = \begin{cases} 0, & \text{if } 0 \leq n(x) < \frac{q}{2q} = 0.5 \\ 1, & \text{if } 0.5 \leq n(x) \leq 1 \end{cases}$$

SO essentially both classifiers perform argmax operation and return 1 for probabilities greater than 0.5, thus we get exactly the classifier we analyzed in class.

Problem 3 - Finding (local) minima of generic functions

(i) Using our assumption, $\exists L \geq 0$ s.t $|f'(z+h) - f'(z)| \leq L^2|z+h-z|$

By dividing the equation with h, we arrive at the definition of the derivative (when h approaches 0) of $f'(z)$ resulting in the second derivative

$$f''(z) = \lim_{h\to 0} \frac{|f'(z+h) - f'(z)|}{|h|} \leq L^2$$

Thus we can conclude $f''(z)$ is bounded.

Using Taylor's Remainder Theorem:

$$f(\bar{x}) = f(x) + f'(x)(\bar{x} - x) + 0.5f''(z)(\bar{x} - x)^2$$

Plugging in $\bar{x} = x - nf'(x)$ gives:

$$f(\bar{x}) = f(x) + f'(x)(x - nf'(x) - x) + 0.5f''(z)(x - nf'(x) - x)^2 = f(x) + f'(x)(-nf'(x)) + 0.5f''(z)(-nf'(x))^2$$

$$f(\bar{x}) = f(x) - nf'(x)^2 + 0.5f''(z)n^2f'(x)^2$$

Because we know $f''(z) \leq L^2$ we get:

$$f(\bar{x}) \leq f(x) - nf'(x)^2 + 0.5L^2n^2f'(x)^2$$

$$f(x) - f(\bar{x}) \geq nf'(x)^2 - 0.5L^2n^2f'(x)^2$$

$$f(x) - f(\bar{x}) \geq nf'(x)^2(1 - 0.5L^2n)$$

- If $f'(x) = 0$ then indeed $f(x) - f(\bar{x}) \geq 0$
- Otherwise for all x, we can pick a small n (explicitly s.t $1 - 0.5L^2n < 0$) to make $f(x) - f(\bar{x}) \geq 0$ only when $f'(x) = 0$

  Solving $(1 - 0.5L^2n) < 0$ gives $n > \frac{2}{L^2}$

  e.g let $n = \frac{4}{L^2}$

  we get $f(\bar{x}) - f(x) \geq \frac{4f'(x)^2}{L^2} \cdot \left(1 - \frac{1}{2} \cdot L^2 \cdot \frac{4}{L^2}\right)$

  $f(x) - f(\bar{x}) \geq \frac{4f'(x)^2}{L^2} \cdot (-1)$

  A positive first multiplied by a negative term yield a negative term. Thus in order for $f(\bar{x}) - f(x) \geq 0$ under those conditions $f'(x) = 0$

(ii)

1. Set $x = x_0$

2. Choose a small positive value for n

3. Compute $f'(x)$

4. Compute $\bar{x} = x - n \cdot f'(x)$

5. if $f(\bar{x}) \leq f(x)$:

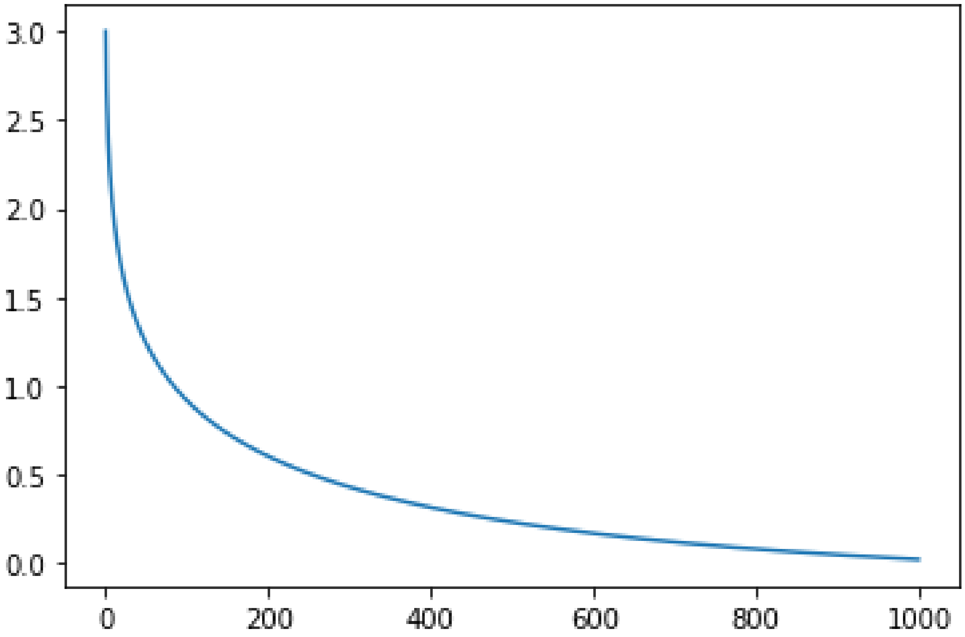
   set $x = \bar{x}$ and repeat from step 3

6. if $f(\bar{x}) > f(x)$:

   reduce the value of n and repeat from step 4

7. repeat steps 3-6 until a stopping criterion is met, such as a small change in f(x) or a maximum number of iterations reached

(iii)



(iv) While this technique allows us to find the local minimum by continuing descending at each iteration, it is not guaranteed to find a global minimum. We can get stuck in a local minimum, especially if the function has multiple local minima, or if the initial value is far from the global minimum. An improvement could be to allow the algorithm at times to escape the local minimum by moving to higher points.

Problem 4 - Exploring the Limits Current Language Models

(i)

$$\mathbb{P}(y|w_{1:n}) = \frac{\mathbb{P}(y, w_{1:n})}{\mathbb{P}(w_{1:n})} = \frac{\mathbb{P}(y) \cdot (\prod_{i=1}^{n} \mathbb{P}(w_i|w_{i-2}w_{i-1}, y))}{\sum_{\tilde{y}} \mathbb{P}(\tilde{y}) \cdot (\prod_{i=1}^{n} \mathbb{P}(w_i|w_{i-2}w_{i-1}, \tilde{y}))}$$

$$\approx \frac{\mathbb{P}(y) \cdot \left(\prod_{i=1}^{n} \frac{C(w_{i-2}w_{i-1}w_i, y)+1}{C(w_{i-2}w_{i-1}, y)+|v|}\right)}{\sum_{\tilde{y}} \mathbb{P}(\tilde{y}) \cdot \left(\prod_{i=1}^{n} \frac{C(w_{i-2}w_{i-1w_i}, \tilde{y})+1}{C(w_{i-2}w_{i-1}, \tilde{y})+|v|}\right)}$$

(ii)(b) OOV bigrams: 0.06664147721972578, on avg 6.5%
OOV trigrams: 0.08259588405203934, on avg 8%
Code submitted on gradescope.

(ii)(c) I have an issue with my code caused by the multiplication of the bigram/trigram probabilities converging to 0, thus I am unable to provide an accuracy rate. I did try to also work with logs instead of probabilities, yet I suspect I have an issue there as well. My assumption is that the accuracy rate would be bigger for the bigram model because the bigram model has a lower OOV rate and has more data to work with (there are more bigrams than trigrams).

(iii)(a)

(iii)(b) N-gram models often fail in this regard because they are only looking at the nearby words, and they generate sentences by predicting each word based on the previous n-1 words, without considering the context beyond those n-1 words.