COMS 4771 Machine Learning (Spring 2023)
Problem Set #1

William Das - whd2108@columbia.edu

February 17, 2023

1 Analyzing Bayes Classifier

(i)

(a)

To compute $P[Y = 1|A, B]$, we integrate over all possible values of C given the join distribution of A,B, and C. Since A, B, and C are independent, we have:

$$P[Y = 1|A, B] = P[Y = 1, A = a, B = b] = P[C < 7 - A - B]$$

We can compute:

$$P[C < 7 - A - B] = \int_0^{7-A-B} e^{-c}\, dc = \left[-e^{-c}\right]_0^{7-a-b} = -e^{-(7-a-b)} - (-e^0) = -e^{a+b-7} + 1$$

Thus:

$$P[Y = 1|A = a, B = b] = -e^{a+b-7} + 1$$

The optimal Bayes classifier is thus:

$$f(\bar{x}) = \arg\max_{y \in Y} P[Y = y|A = a, B = b]$$

or

$$f(a, b) = \begin{cases} 1 & \text{if } -e^{a+b-7} + 1 > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

If we let $Z = A + B$, where $P[Y = 1|Z = z] = -e^{z-7} + 1$, we can define the misclassification rate, or Bayes error rate, for this classifier as sum of the probabilities that a predicted label does not match the true label:

$$P[\hat{Y} \neq y] = \sum_{\hat{y} \in Y} \int_h^k P[Z = z] \cdot P[\hat{Y} \neq \hat{y}|Z = z]\, dz$$

where in each summation, the interval $[h, k]$ represents the region which classifies class $\hat{y}$ as true.

To get the bounds for each interval, we solve for z in $P[Y = 1|Z = z] > \frac{1}{2}$ to obtain the interval, or region, in which the predictor classifies Z as 1:

$$P[Y = 1|Z = z] = -e^{z-7} + 1 > \frac{1}{2}$$

$$e^{z-7} < \frac{1}{2}$$

$$z < 7 - ln(2)$$

For the sum of exponential distributions Z, 0 to $7 - ln(2)$ is the region of classification for 1, and $7 - ln(2)$ to 7 is the region of classification for 0. Our sum of integrals becomes:

$$\int_0^{7-ln(2)} P[Z] \cdot (1 - (-e^{z-7} + 1))\, dz + \int_{7-ln(2)}^{7} P[Z] \cdot (1 - e^{z-7})\, dz$$

$$= \int_0^{7-ln(2)} P[Z] \cdot e^{z-7}\, dz + \int_{7-ln(2)}^{7} P[Z] \cdot (1 - e^{z-7})\, dz$$

To obtain $P[Z]$, where Z is the sum of two exponential random variables A and B, we can convolve the pdfs of A and B:

$$f_Z(z) = f_A(z) \cdot f_B(z) = \int_{-\infty}^{\infty} f_A(a) \cdot f_B(z - a)\, da = \int_0^z e^{-a} \cdot e^{-(z-a)}\, da$$

$$\int_0^z e^{-a} \cdot e^{-(z-a)}\, da = \int_0^z e^{-z}\, da = e^{-z} \int_0^z 1\, da = e^{-z}\, [a]_0^z = ze^{-z}$$

Plugging this into the Bayes error rate, we have:

$$\int_0^{7-ln(2)} ze^{-z} \cdot e^{z-7}\, dz + \int_{7-ln(2)}^{7} ze^{-z} \cdot (1 - e^{z-7})\, dz$$

$$= \int_0^{7-ln(2)} ze^{-7}\, dz + \int_{7-ln(2)}^{7} ze^{-z}\, dz - \int_{7-ln(2)}^{7} ze^{-7}\, dz$$

$$= e^{-7} \int_0^{7-ln(2)} z\, dz + \int_{7-ln(2)}^{7} ze^{-z}\, dz - e^{-7} \int_{7-ln(2)}^{7} z\, dz$$

An integration of parts on ze^{-z} yields:

$$\int ze^{-z}\, dz = uv - \int v\, du$$

where

$$u = z$$

$$dv = e^{-z}dz$$
$$v = \int e^{-z}\,dz = -e^{-z}$$
$$du = dz$$

We have:

$$\int ze^{-z}\,dz = uv - \int v\,dv = -ze^{-z} - \int -e^{-z}\,dz = -ze^{-z} - e^{-z} = -e^{-z}(z+1)$$

Plugging this into the error rate:

$$e^{-7}\int_0^{7-ln(2)} z\,dz + \int_{7-ln(2)}^{7} ze^{-z}\,dz - e^{-7}\int_{7-ln(2)}^{7} z\,dz$$
$$= e^{-7}\left[\frac{z^2}{2}\right]_0^{7-ln(2)} + \left[-e^{-z}(z+1)\right]_{7-ln(2)}^{7} - e^{-7}\left[\frac{z^2}{2}\right]_{7-ln(2)}^{7}$$
$$\approx .01996$$

(b)

To compute $P[Y = 1|A = a]$, we can set $Z = B + C$. Using the integration by parts from (a), we have:

$$P[Y = 1|A] = P[B + C < 7 - A] = P[Z < 7 - A] = \int_0^{7-a} ze^{-z}\,dz = [-ze^{-z} + e^{-z}]_0^{7-a}$$

After an integration by parts of ze^{-z} as shown in (a), we have:

$$\int_0^{7-a} ze^{-z}\,dz = \left[-ze^{-z} + e^{-z}\right]_0^{7-a} = \left[-e^{-z}(z+1)\right]_0^{7-a} = [-e^{-(7-a)}(7-a+1)] - [-1\cdot 1] = -e^{a-7}(8-a)+1$$

We have:

$$P[Y = 1|A = a] = -e^{a-7}(8-a) + 1$$

The optimal Bayes classifier is thus:

$$f(\bar{x}) = \arg\max_{y \in Y} P[Y = 1|A = a]$$

or

$$f(a) = \begin{cases} 1 & \text{if } -e^{a-7}(8-a) + 1 > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

For the Bayes error rate, we can use the same equation in (a). To get the bounds from [0, 7], we calculate:

$$-e^{a-7}(8-a) > \frac{1}{2}$$

Graphing this function, we find that $a \leq 5.32$ or $a \geq 7.77$. We can take the interval from [0, 5.32] as the region of classification for class 1, and [5.32, 7] as the region of classification for class 0. Our Bayes error rate now is:

$$\int_0^{5.32} P[A] \cdot (1 - (-e^{a-7}(8-a)+1))\, da + \int_{5.32}^{7} P[A] \cdot (-e^{a-7}(8-a)+1)\, da$$

$P[A]$ is the pdf of A, or e^{-a}. Substituting:

$$\int_0^{5.32} P[A] \cdot (1 - (-e^{a-7}(8-a)+1))\, da + \int_{5.32}^{7} P[A] \cdot (-e^{a-7}(8-a)+1)\, da$$

$$= \int_0^{5.32} e^{-a}e^{a-7}(8-a)\, da + \int_{5.32}^{7} e^{-a}(-e^{a-7}(8-a)+1)\, da$$

$$= e^{-7}\int_0^{5.32} 8-a\, da + \int_{5.32}^{7} -e^{-7}(8-a) + e^{-a}\, da$$

$$= e^{-7}\int_0^{5.32} 8-a\, da - e^{-7}\int_{5.32}^{7} 8-a\, da + \int_{5.32}^{7} e^{-a}\, da$$

$$= e^{-7}\left[8a - \frac{a^2}{2}\right]_0^{5.32} - e^{-7}\left[8a - \frac{a^2}{2}\right]_{5.32}^{7} + \left[-e^{-a}\right]_{5.32}^{7}$$

$$\approx .027$$

(c)

We can represent the sum of 3 exponential, independent random variables as $Z = A+B+C$. We can convolve A+B again with C to obtain the cdf of this distribution:

$$f_Z(z) = f_{A+B}(z) \cdot f_C(z) = \int_{-\infty}^{\infty} f_{A+B}(x) \cdot f_C(z-x)\, dx = \int_0^z xe^{-x} \cdot e^{-(z-x)}\, dx = \int_0^z xe^{-z}\, dx$$

$$= e^{-z}\int_0^z x\, dx = e^{-z}\left[\frac{x^2}{2}\right]_0^z = \frac{z^2e^{-z}}{2}$$

Now we have:

$$P[Y=1|Z=z] = \int_0^7 \frac{z^2e^{-z}}{2}\, dz = \frac{1}{2}\int_0^7 z^2e^{-z}\, dz$$

Integrating by parts:

$$\int_0^7 z^2e^{-z}\, dz = uv - \int v\, du$$

where

$$u = z^2$$

$$dv = e^{-z}dz$$

$$v = \int e^{-z}\, dz = -e^{-z}$$

$$du = 2z\,dz$$

We have:

$$\int_0^7 z^2 e^{-z}\,dz = uv - \int v\,dv = -z^2e^{-z} - \int -2ze^{-z}\,dz$$

From (a):

$$\int -2ze^{-z}\,dz = -2\int ze^{-z}\,dz = -2(-e^{-z})(z+1) = 2e^{-z}(z+1)$$

The completed integral now is:

$$\int_0^7 z^2 e^{-z}\,dz = [-z^2e^{-z} - 2e^{-z}(z+1)] = \left[-e^{-z}(z^2+2z+2)\right]_0^7$$

$$\frac{1}{2}\left[-(z^2+2z+2)e^{-z}\right]_0^7 = \frac{1}{2}[-(49+14+2)e^{-7} - (-2)] = \frac{1}{2}[-65e^{-7}+2] = \frac{-65e^{-7}+2}{2} = .97$$

The optimal Bayes classifier is thus:

$$f(\bar{x}) = \arg\max_{y \in Y} P[Y = 1|Z] = 1$$

This Bayes classifier will always be 1 as $P[Y = 1|Z] = .97 > .5$. As the Bayes classifier always predicts 1 with probability .97, the classification error will simply be:

$$1 - ((.97)(1) - (.03)(0)) = .03$$

(ii)

If we set $Z = A + B$, we obtain $P[Y = 1|Z = z] = P[C < 7 - Z]$. If we examine a uniform distribution $C \sim U[-n, n]$ which exists on the interval from -n to n, we can observe the value of $P[Y = 1|Z = z]$ for large values of n, and examine the Bayes error rate.

Using the pdf of a uniform distribution, we can compute $P[Y = 1|Z = z]$ as:

$$P[Y = 1|Z = z] = P[C < 7-Z] = \int_{-n}^{7-z} \frac{1}{n-(-n)}\,dz = \int_{-n}^{7-z} \frac{1}{2n}\,dz = \left[\frac{z}{2n}\right]_{-n}^{7-z} = \frac{7-z+n}{2n}$$

We can notice that:

$$\lim_{n\to\infty} \frac{7-z+n}{2n} = \frac{1}{2}$$

since n dominates in the limit, specifically $\frac{n}{2n}$.

If the $P[X = 1|Z]$ approaches 0, then $P[X = 0|Z]$ will also simultaneously approach 0. If both probabilities are as close to each other as possible, and both approaching $\frac{1}{2}$, then we have as the Bayes error rate:

$$1 - (\frac{1}{2}\cdot\frac{1}{2} + \frac{1}{2}\cdot\frac{1}{2}) = \frac{1}{2}$$

2 Classification with Asymmetric Costs

(i)

In medical diagnosis, or binary classification of diseases, we care about the confidence of decisions outputted by a classifier, as false negatives or false positives can adversely impact a patient's well-being—in the case of a false negative, a patient may not receive needed treatment, and in the case of a false positive, a patient may receive unnecessary treatment. If the model is uncertain of a decision, then this can be taken into account to prevent outputting a false diagnosis. Moreover, the model can take into account different, asymmetric losses based on the true and predicted values, which allows us to weight certain losses, such as false positives or false negatives, higher.

(ii)

Given

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

$$\eta(x) = Pr[Y = 1 | X = x]$$

$$r < \frac{pq}{p+q}$$

We want to show that for any other classifier $g : X \to -1, 0, 1$ that:

$$\mathbb{E}_{x,y}\ell(f^*(x), y) \leq \mathbb{E}_{x,y}\ell(g(x), y)$$

Or that:

$$\mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) \geq 0$$

First we can adapt the indicator function defined in class, which takes a boolean x:

$$\mathbf{1}(x) = \begin{cases} 0 & \text{if } x \text{ is false} \\ 1 & \text{if } x \text{ is true} \end{cases}$$

To start, we can represent the $\mathbb{E}_{x,y}\ell(f^*(x), y))$ as:

$$p \cdot \eta(x)\mathbf{1}[f^*(x) = 0] + q \cdot (1 - \eta(x))\mathbf{1}[f^*(x) = 1] + r \cdot \mathbf{1}[f^*(x) = -1]$$

Applying the same formula to $g(x)$ and subtracting from $\mathbb{E}_{x,y}\ell(f^*(x), y))$, we can distribute the differences accordingly, as shown in class. The equation below can be denoted as **Equation 1**:

$$\begin{aligned} \mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) = & p \cdot \eta(x)[\mathbf{1}[g(x) = 0] - \mathbf{1}[f^*(x) = 0]] + q \cdot (1 - \eta(x))[\mathbf{1}[g(x) = 1] \\ & - \mathbf{1}[f^*(x) = 1]] + r \cdot [\mathbf{1}[g(x) = -1] - \mathbf{1}[f^*(x) = -1]] \end{aligned}$$

We can conduct a case analysis on the values that $g(x)$ and $f^*(x)$ can take to show that, for all possible values and the corresponding expected value of the classifier loss, the loss of the the optimal Bayes classifier will be less than or equal to that of any binary classifier, or that $\mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) \geq 0$.

The first three cases are trivial, in which $g(x) = f^*(x) = 0$ or 1 or -1. In these cases, the difference of the indicator functions in each coefficient evaluates to 0, so the sum of terms on the RHS of Equation 1 will be 0: $\mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) = 0 \geq 0$.

The next cases are:

Case 4: $g(x) = 0, f^*(x) = 1$

The RHS of Equation 1 becomes:

$$p\eta(x) \cdot 1 + q(1 - \eta(x)) \cdot -1 = p\eta(x) - q + q\eta(x) = \eta(x)(p + q) - q$$

When $f^*(x) = 1$, $\eta(x) \geq 1 - \frac{r}{q}$. Substituting:

$$\eta(x)(p + q) - q \geq (1 - \frac{r}{q})(p + q) - q$$

Since $r < \frac{pq}{p+q}$, we have:

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

Therefore:

$$\eta(x)(p + q) - q \geq 0$$

Case 5: $g(x) = 0, f^*(x) = -1$

The RHS of Equation 1 becomes:

$$p\eta(x) \cdot 1 + r \cdot -1 = p\eta(x) - r$$

When $f^*(x) = -1$, $\eta(x) > \frac{r}{q}$. So:

$$p\eta(x) - r > p \cdot \frac{r}{p} - r = r - r = 0 \geq 0$$

Therefore:

$$p\eta(x) - r \geq 0$$

Case 6: $g(x) = 1, f^*(x) = 0$

The RHS of Equation 1 becomes:

$$p\eta(x) \cdot -1 + q(1 - \eta(x)) = -p\eta(x) + q - q\eta(x) = -\eta(x)(p+q) + q$$

When $f^*(x) = 0$, $\eta(x) \leq \frac{r}{p}$. Substituting:

$$-\eta(x)(p+q) + q \geq -\frac{r}{p}(p+q) + q$$

Using constraints on r:

$$-\frac{r}{p}(p+q) + q > -\frac{\frac{pq}{p+q}}{p}(p+q) + q = -\frac{q}{p+q}(p+q) + q = -q + q = 0 \geq 0$$

Therefore:

$$-\eta(x)(p+q) + q \geq 0$$

Case 7: $g(x) = 1, f^*(x) = -1$

The RHS of Equation 1 becomes:

$$q(1 - \eta(x)) + r \cdot -1 = -q\eta(x) + q - r$$

When $f^*(x) = -1$, $\eta(x) < 1 - \frac{r}{q}$. Substituting:

$$-q\eta(x) + q - r > -q(1 - \frac{r}{q}) + q - r = -q + r + q - r = 0 \geq 0$$

Therefore:

$$-q\eta(x) + q - r \geq 0$$

Case 8: $g(x) = -1, f^*(x) = 0$

The RHS of Equation 1 becomes:

$$p\eta(x) \cdot -1 + r = -p\eta(x) + r$$

When $f^*(x) = 0$, $\eta(x) \leq \frac{r}{q}$. Substituting:

$$-p\eta(x) + r \geq -p((\frac{r}{p}) + r = -r + r = 0 \geq 0$$

Therefore:

$$-p\eta(x) + r \geq 0$$

Case 9: $g(x) = -1, f^*(x) = 1$

The RHS of Equation 1 becomes:

$$q(1 - \eta(x)) \cdot -1 + r = q\eta(x) - q + r$$

When $f^*(x) = 1$, $\eta(x) \geq 1 - \frac{r}{q}$. Substituting:

$$q\eta(x) - q + r \geq q(1 - \frac{r}{q}) - q + r = q - r - q + r = 0 \geq 0$$

Therefore:

$$q\eta(x) - q + r \geq 0$$

We've shown that for all possible cases of values for $g(x)$ and $f^*(x)$:

$$\mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) \geq 0$$

As such, the inequality must hold true, and f^* is the Bayes classifier for the model of classification posed in the problem.

(iii)

Given

$$f^*(x) = \begin{cases} 0 & \text{if } 0 \leq \eta(x) \leq \frac{q}{p+q} \\ 1 & \text{if } \frac{q}{p+q} \leq \eta(x) \leq 1 \end{cases}$$

$$\eta(x) = Pr[Y = 1|X = x]$$

$$r \geq \frac{pq}{p+q}$$

We want to show that for any other classifier $g : X \to -1, 0, 1$ that:

$$\mathbb{E}_{x,y}\ell(f^*(x), y) \leq \mathbb{E}_{x,y}\ell(g(x), y)$$

Or that:

$$\mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) \geq 0$$

To start, we can modify the equation from (ii), and remove the indicator function where $f^*(x) = -1$, denoted as **Equation 2**:

$$\mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) = p \cdot \eta(x)[\mathbf{1}[g(x) = 0] - \mathbf{1}[f^*(x) = 0]] + q \cdot (1 - \eta(x))[\mathbf{1}[g(x) = 1] - \mathbf{1}[f^*(x) = 1]] + r \cdot [\mathbf{1}[g(x) = -1]]$$

We can conduct a similar case analysis as in (ii), and show that the inequality in Equation 2 holds for all cases of $g(x)$ and $f^*(x)$ values.

The first two trivial cases are in which $g(x) = f^*(x) = 0$ or 1, in which the difference

in expected values will equal $0 \geq 0$, as seen in (ii).

The next cases are:

Case 3: $g(x) = 0, f^*(x) = 1$

The RHS of Equation 2 becomes:

$$p\eta(x) + q(1 - \eta(x)) \cdot -1 = p\eta(x) - q + q\eta(x) = \eta(x)(p + q) - q$$

When $f^*(x) = 1$, $\eta(x) \geq \frac{q}{p+q}$. Substituting:

$$\eta(x)(p + q) - q \geq \frac{q}{p + q}(p + q) - q = 0 \geq 0$$

Therefore:

$$\eta(x)(p + q) - q \geq 0$$

Case 4: $g(x) = 1, f^*(x) = 0$

The RHS of Equation 2 becomes:

$$p\eta(x) \cdot -1 + q(1 - \eta(x)) = -p\eta(x) + q - q\eta(x) = -\eta(x)(p + q) + q$$

When $f^*(x) = 0$, $\eta(x) \leq \frac{q}{p+q}$. Substituting:

$$-\eta(x)(p + q) + q \geq -\frac{q}{p + q}(p + q) + q = -q + q = 0 \geq 0$$

Therefore:

$$-\eta(x)(p + q) + q \geq 0$$

Case 5: $g(x) = -1, f^*(x) = 0$

The RHS of Equation 2 becomes:

$$-p\eta(x) + r$$

When $f^*(x) = 0$, $\eta(x) \leq \frac{q}{p+q}$. Substituting and using the constraints on r:

$$-p\eta(x) + r \geq -\frac{pq}{p + q} + r \geq -\frac{pq}{p + q} + \frac{pq}{p + q} = 0 \geq 0$$

Therefore:

$$-p\eta(x) + r \geq 0$$

Case 6: $g(x) = -1, f^*(x) = 1$

The RHS of Equation 2 becomes:

$$-q(1 - \eta(x)) + r = q\eta(x) - q + r$$

When $f^*(x) = 1$, $\eta(x) \geq \frac{q}{p+q}$. Substituting, and using constraints on r, we have:

$$q\eta(x) - q + r \geq q \cdot \frac{q}{p+q} - q + \frac{pq}{p+q} = \frac{q^2 - q(p+q) + pq}{p+q} = 0 \geq 0$$

Therefore:

$$q\eta(x) - q + r \geq 0$$

We've shown that for all possible cases of values for $g(x)$ and $f^*(x)$:

$$\mathbb{E}_{x,y}\ell(g(x), y) - \mathbb{E}_{x,y}\ell(f^*(x), y) \geq 0$$

As such, the inequality must hold true, and f^* is the Bayes classifier for the model of classification posed in the problem.

(iv)

Given $p = q$ and $r > \frac{p}{2}$:

For (ii), if $r > \frac{p}{2}$, then

$$\frac{r}{p} > \frac{\frac{p}{2}}{p} = \frac{1}{2}$$

and

$$1 - \frac{r}{p} < 1 - \frac{\frac{p}{2}}{p} = \frac{1}{2}$$

Substituting these values in (ii), the $f^*(x)$ is now:

$$f^*(x) = \begin{cases} 0 & \text{if } 0 \leq \eta(x) \leq \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} < \eta(x) < \frac{1}{2} \\ 1 & \text{if } 1 - \frac{1}{2} \leq \eta(x) \leq 1 \end{cases}$$

$f^*(x)$ can never be -1, as $\eta(x)$ is bounded by $(\frac{1}{2}, \frac{1}{2})$. As such, $f^*(x)$ is reduced to:

$$f^*(x) = \begin{cases} 0 & \text{if } 0 \leq \eta(x) \leq \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq \eta(x) \leq 1 \end{cases}$$

This is the same definition of the Bayes classifier we derived in class: $f(\hat{(}x)) = \arg\max P[Y = y|X = \hat{x}]$.

Moreover, in (iii), substituting values of p and q into the definition of $f^*(x)$ will yield the same Bayes classifier:

$$f^*(x) = \begin{cases} 0 & \text{if } 0 \leq \eta(x) \leq \frac{1}{2} \\ 1 & \text{if } \frac{1}{2} \leq \eta(x) \leq 1 \end{cases}$$

3 Finding (local) minima of generic functions

(i)

Assuming that there exists $L \geq 0$ such that for all $a, b \in \mathbb{R}$, $|f'(a)f'(b)| \leq L^2|a-b|$, prove the following statement:

For any $x \in \mathbb{R}$, there exists some $\eta > 0$, such that if $\bar{x} := x - \eta f'(x)$, then $f(\bar{x}) \leq f(x)$, with equality if and only if $f'(x) = 0$.

Proof: To show that f has a bounded second derivative, we can first obtain a second derivative by taking the limit of $a \to b$, for $a, b \in \mathbb{R}$:

$$|f''(z)| = \lim_{a \to b} \frac{|f'(a) - f'(b)|}{|b-a|}$$

Using our assumption, we know:

$$\frac{|f'(a) - f'(b)|}{|a-b|} \leq L^2$$

Taking the limits of both sides of the inequality:

$$|f''(z)| = \lim_{a \to b} \frac{|f'(a) - f'(b)|}{|b-a|} \leq \lim_{a \to b} L^2$$

L^2 is a constant, so we have:

$$|f''(z)| = \lim_{a \to b} \frac{|f'(a) - f'(b)|}{|b-a|} \leq L^2$$

$$|f''(z)| \leq L^2$$

This implies that f has a bounded second derivative for all z, bounded by L^2.

To apply Taylor's Remainder Theorem, we can find a $z \in (x, \overline{x})$ such that:

$$f(\bar{x}) = f(x) + f'(x)(\bar{x} - x) + \frac{1}{2} f''(z)(\bar{x} - x)^2$$

We know from our assumptions that $\bar{x} := x - \eta f'(x)$. So then: $\bar{x} - x := -\eta f'(x)$.

Substituting for $\bar{x} - x$:

$$f(\bar{x}) = f(x) + f'(x)(-\eta f'(x)) + \frac{1}{2} f''(z)(-\eta f'(x))^2$$

$$f(\bar{x}) = f(x) - \eta f'(x)^2 + \frac{1}{2} f''(z)\eta^2 f'(x)^2$$

$$f(\bar{x}) - f(x) = -\eta f'(x)^2 + \frac{1}{2} f''(z)\eta^2 f'(x)^2$$

$$f(\bar{x}) - f(x) = -\eta f'(x)^2(1 - \frac{1}{2}\eta f''(z))$$

Multiplying both sides by -1:

$$f(x) - f(\bar{x}) = \eta f'(x)^2(1 - \frac{1}{2}\eta f''(z))$$

We want to show that $f(\bar{x}) \leq f(x)$ for some $\eta > 0$. On the RHS, $\eta f'(x)^2$ will always be greater than or equal to 0, so we want the expression $1 - \frac{1}{2}\eta f''(z) > 0$. Since we know $|f''(z)| \leq L^2$, we have:

$$1 - \frac{1}{2}\eta L^2 > 0$$

Rearranging, we must have:

$$\eta < \frac{2}{L^2}$$

Thus, given the constraint on L^2, we can obtain a positive $0 < \eta < \frac{2}{L^2}$ that satisfies the inequality $f(\bar{x}) \leq f(x)$. Given that the bounds on η is satisfied, then $f(x) - f(\bar{x})$ can only be 0 if $f'(x)$ is 0 in the product of terms on the RHS.

(ii)

We can devise an iterative algorithm as follows:

1. Choose an initial value x_0.
2. While the derivative of the function is not close to zero at x_0, do the following:
   (a) Compute $x_1 = x_0 - \eta f'(x_0)$, where $f'(x)$ is the gradient of f.
   (b) If $f(x_1) < f(x_0)$, set $x_0 = x_1$ and repeat step 2.
   (c) Keep track of the x points, and corresponding $f(x)$ values.

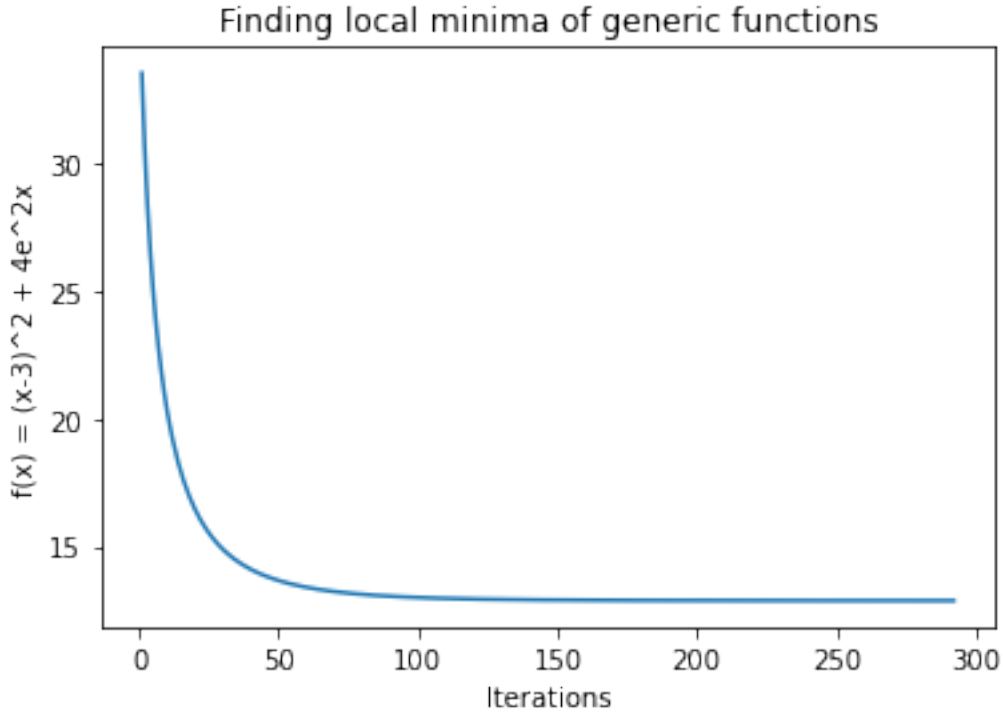
(iii)

A plot of of iterations against local minima at each step in iteration of $f(x)$ is shown in Figure 1. The local minima converges at around 300 iterations, and the estimated value is $f(-0.123) = 12.8808$.

(iv)

This technique is only useful for finding local minima since the algorithm can start at an arbitrary start point, and repeatedly estimate local minima, terminating when the derivative of the function at the current value of x is close to 0. This will lead to suboptimal local minima estimates, as the algorithm terminates when it converges to a value close to zero, and relies on the gradient—which looks at points in the immediate neighborhood of a current point—at each step to estimate the next minimal point and stop accordingly.

Figure 1: Plot of of iterations against f(x) for estimating local minima.



Some possible improvements are initializing different x_0 across a specified range over the domain of f, and comparing the local minima obtained from each initial point to obtain a global minima that encompasses a larger range of x values.

4 Exploring the Limits of Current Language Models

(i)

If we define $C(w_{1:n}, y)$ as the number of occurrences of a bigram, trigram, or n-gram of length n in class y, we can write $P(y|w_{i:n})$ as:

$$P(y|w_{i:n}) = \frac{P(y)P(w_{i:n}|y)}{\sum_{\tilde{y}} P(\tilde{y})P(w_{i:n}|\tilde{y})}$$

where

$$P(w_{1:n}|y) = \prod_{i=1}^{n} \frac{C(w_{i-2}w_{i-1}w, y) + 1}{C(w_{i-2}w_{w-1}, y) + |V|}$$

(ii)

Code is attached in n_gram_classifier.py, which requires that the humvgpt folder is in the same directory to run.

I removed all punctuation except for ,.?", and evaluated model performance after training using the sum and difference of log probabilities corresponding to conditional and class priors, comparing log likelihoods between the two classes.

(b)

Bigram OOV Rate: 0.0981
Trigram OOV Rate: 0.3542

(c)

The bigram model had an accuracy of 96.02%, and the trigram model had an accuracy of 95.16%. Overall, the accuracy metrics are very strong, with the bigram model performing marginally better than the trigram model. While the bigram model only performs marginally better than the trigram model, given the OOV rates present in the test set, theoretically the bigram model should perform better than the trigram model.

Higher OOV rates should lead to poorer classification performance, as the classifier is unable to estimate the probability of an unseen ngram, and resorts to a smoothing operation to account for this in order to prevent zero probabilities. The bigram model, with a lower OOV rate, is able to estimate the probabilities of its associated bigrams more accurately, and incorporate more bigrams to take into account when classifying.

In the probability estimations, we use additive, or Laplacian smoothing, to account for tokens that aren't present—this can lead to poor performance, as all unseen ngrams are assigned the same probability, contingent on the vocabulary size, and this assigned probability can take away the model's emphasis on other present ngrams. As a result, since the trigrams

have a higher OOV rate, coupled with Laplacian smoothing to account for this, the trigram model should theoretically perform worse than the bigram model.

(iii)

(a)

Below is example output of sentences generated using the bigram and trigram models. The GPT Trigram model seems to generate the best sentences overall. Overall, the GPT corpous seems to relatively generate more flowing sentences.

As temperature increases, the sentences seem to become more complex and marginally cohesive. The trigram model relatively generates more complex sentences than the bigram model—in the bigram model, especially towards lower temperatures (T=20), a cycle of repeated tokens occurs. This is likely due to the fact that the probabilities are magnified exponentially when the temperature decreases, resulting in probabilities that are consistently favored towards certain phrases—this leads to more repeated phrases that the model is comfortable with. With higher temperatures, the probabilities are more dispersed, allowing for more variety in the weighted random sampling of next tokens, but may lead to more incoherent and nonsensical sentences.

Sentence Generation (GPT, Bigrams, T=50):
fair and the same way to the same way to the same way to the same way to the united states ,

Sentence Generation (Hum, Bigrams, T=50): magnets and the same thing . the same thing . the same way to the same way to the same thing .

Sentence Generation (GPT, Trigrams, T=50):
homebrew is software that claims to have a lot of money you would need to be able to slow down or stop and

Sentence Generation (Hum, Trigrams, T=50):
visiting a website is nt really , due to the real world . if you have to be a good idea to keep

Sentence Generation (GPT, Bigrams, T=50):
facing a good idea to the same way to the same way to the same way to the united states , and

Sentence Generation (GPT, Bigrams, T=500):
net neutrality , and the same frequency of the mistake to the same language developed and the same time . the same

Sentence Generation (Hum, Bigrams, T=500):
linkedin profile . the same direction . the same thing about it s not , and the same as a lot of

Sentence Generation (GPT, Trigrams, T=500):
contestants on game developers and publishers to sell the stock price will depend on a trade is executed at the federal reserve , the

Sentence Generation (Hum, Trigrams, T=500):
albums are , but they do nt have more money and pay the fine print items on the back of the fire department

Sentence Generation (GPT, Bigrams, T=1000):
knowing how much heavier and the quickest route and the price that the price if you can be aware of the other

Sentence Generation (Hum, Bigrams, T=1000):
read this a lot of the same way to be able to the camera takes the same amount of the river .

Sentence Generation (GPT, Trigrams, T=1000):
instrumental convergence can be convenient for customers . its important to note that there may be certain restrictions in place for the government

Sentence Generation (Hum, Trigrams, T=1000):
accents are british accents . a one way to start fixing problems . tldr he offered runaways something their homes . the state

Sentence Generation (GPT, Bigrams, T=20):
chuck and the same way to the same way to the same way to the same way to the same way to

Sentence Generation (Hum, Bigrams, T=20):
global warming the same way to the same way to the same way to the same way to the same thing .

Sentence Generation (GPT, Trigrams, T=20):
paedophilia is a type of propulsion systems , as well as the stock market is a type of person who is interested in

Sentence Generation (Hum, Trigrams, T=20):
shops in most states have a lot of people who are rh can not be able to take a look at the same way

(b)

Bigram and trigram models only consider the prior two or three tokens, failing to capture long-range context as transformers that employ "attention mechanisms" do. As a result, as shown in the generated examples, the sentences come out to be incoherent as the next token

does not take into account all the preceding tokens, but rather the last two or three tokens.