Conditional Probability and Bayes Identity:

$$\begin{aligned} Pr\{A|B\} &= \frac{Pr\{B \;\&\; A\}}{Pr\{B\}} \\ Pr\{A|B\} &= \frac{Pr\{B|A\}\;\; Pr\{A\}}{Pr\{B\}} \end{aligned}$$

Derivatives and logs:

$$\begin{aligned} \frac{d\,(x^n)}{dx} &= nx^{n-1} \\ \frac{d\,(e^x)}{dx} &= e^x \\ \frac{d\,(\log\; x)}{dx} &= \frac{1}{x} \\ \frac{\partial\,[\,f(x,y)g(x,y)]}{\partial x} &= f(x,y)\frac{\partial g(x,y)}{\partial x} + \frac{\partial f(x,y)}{\partial x}g(x,y) \end{aligned}$$

$$\begin{aligned} \log(xy) &= \log x + \log y \\ \log(1/x) &= -\log x \\ \log(x^a) &= a\log x \\ \log(e) &= 1 \\ \rightarrow \log(e^x) &= x \end{aligned}$$

Basic chain rule:

$$\frac{d}{dx}f\,[g\,(x)] = \frac{df}{dg}\,\frac{dg}{dx}$$

$$\text{e.g.,} \qquad \frac{d}{dr}e^{-r^2/2} = -re^{-r^2/2}$$

Name: _____ ID#
(print) (last) (first)

1. Consider cardiovascular healthscores based on cholesterol levels:

Low risk (L): cholesterol level ≤ 200 mg/dL

Moderate risk (M): 200 < cholesterol level ≤ 240 mg/dL

High risk (H): cholesterol level > 240 mg/dL

(a) A researcher assumes (based on domain knowledge) that 'L' is twice as probable as 'M', i.e., if $Pr\{M\} = p$ then $Pr\{L\} = 2p$. They sample three members of a small town in order to estimate p, and obtain three independent scores: 'L', 'M', and 'H'. Write the log likelihood function for this model and these data. Given the model, the data, and the likelihood function, what is the best estimate of p?

(b) What would be the answer if the model kept the three categories but did not assume that 'L' is twice as probable as 'M'. That is, if the model does not assume anything in particular about the three probabilities such that $Pr\{M\} = p$ and $Pr\{L\} = q$ (and of course $p + q \leq 1$)? (Note the this model has two parameters that need to be estimated, rather than one).

a) $L = p \cdot 2p \cdot (1-3p)$

$l = \log(p) + \log(2p) + \log(1-3p)$

$\frac{dl}{dp} = \frac{1}{p} + \frac{2}{2p} + \frac{-3}{1-3p} = 0$

$\frac{2}{p} = \frac{3}{1-3p}$

$2 - 6p = 3p$

$2 = 9p$

$p = \frac{2}{9}$

Name: _____ ID#
(print) (last) (first)

b) $L = p \cdot q \cdot (1-p-q)$

$\ell = \log(p) + \log(q) + \log(1-p-q)$

$\frac{d\ell}{dp} = \frac{1}{p} - \frac{1}{1-p-q} = 0$ (eq 1)

$\frac{d\ell}{dq} = \frac{1}{q} - \frac{1}{1-p-q} = 0$ (eq 2)

$\frac{1}{p} - \frac{1}{1-p-q} = \frac{1}{q} - \frac{1}{1-p-q}$ (eq 1 = eq 2)

$\frac{1}{p} = \frac{1}{q} \rightarrow p = q$

$\frac{1}{p} - \frac{1}{1-2p} = 0$ (eq 1, substitute $q=p$)

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

$1-2p = p$

$1 = 3p$

$p = \frac{1}{3}$ $q = \frac{1}{3}$

Name: _____ ID#
(print) (last) (first)

2. (a) In a given week during (a severe) flu season, 1% of the population of Illinois has the flu. A diagnostic test for flu has a false positive rate of 2%, i.e., a probability $p = 0.02$ to indicate 'positive' when the subject being tested does **not** have the flu. The same test has a false negative rate of 1% – a probability of $p = 0.01$ of indicating 'negative' when the subject has flu. If a randomly selected individual tests positive, what is the probability that they actually have the disease?

(b) Suppose that 40% of your email can be labeled 'interesting' and 30% is 'important'. Moreover, 25% of the interesting email is actually important. Draw a Venn diagram that corresponds to these observations and label each of the three distinct regions in the diagram with the corresponding probability. In addition, find the conditional probability that a randomly selected email is interesting given that it is important.

a) $P(\text{disease}|\text{pos}) = \dfrac{P(\text{pos}|\text{dis}) \cdot P(\text{dis})}{P(\text{pos})}$

$P(\text{pos}|\text{dis}) = 1 - FN = 1 - 0.01 = 0.99$

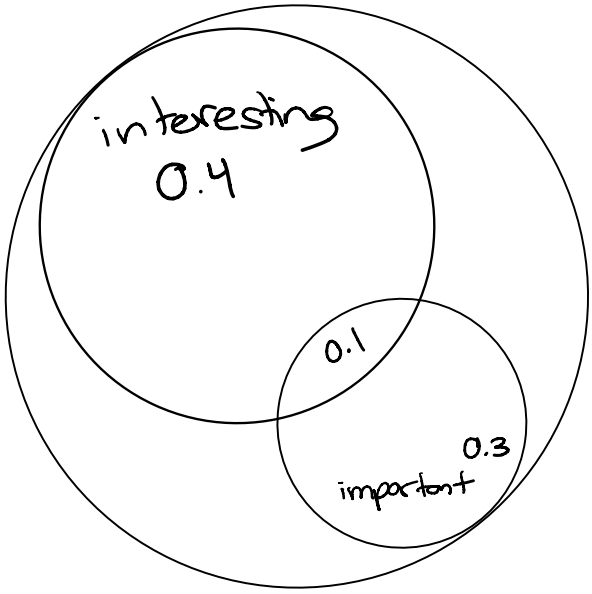
$P(\text{dis}) = 0.01$

$P(\text{pos}) = \underbrace{0.99(0.01)}_{\text{has disease}} + \underbrace{0.02(0.99)}_{\text{false positive}} = 0.0297$

$P(D|P) = \dfrac{0.99 \cdot 0.01}{0.0297} = \dfrac{1}{3}$

Name: _____ ID#
(print) (last) (first)

b)



$0.25(0.4) = 0.1$

$$P(\text{Int} \mid \text{Imp}) = \frac{P(\text{Imp} \mid \text{Int})\,P(\text{Int})}{P(\text{Imp})} = \frac{0.25 \cdot 0.4}{0.3}$$

$$= \boxed{\frac{1}{3}}$$

Name: _____ ID#
(print) (last) (first)

3. For a given data point, the loss function is

$$\mathcal{L}(\beta_1, \beta_2) = e^{\beta_1^2 - \beta_2^2 + 4\beta_2}$$

Suppose that the current values of the model parameters are

$$\beta_1^{(t)} = 1 \quad \text{and} \quad \beta_2^{(t)} = 2$$

(a) Manually perform **one iteration** of the gradient descent algorithm to minimize $f(\beta_1, \beta_2)$, i.e., finding $\beta_i^{(t+1)}$ when the training data is that same single point. Use a (learning) rate of $\gamma = 0.5e^{-5}$.
Is $f\left(\beta_1^{(t+1)}, \beta_2^{(t+1)}\right) < f\left(\beta_1^{(t)}, \beta_2^{(t)}\right)$? If not, what might be the problem?

(b) Repeat the calculation from (a), but this time use a learning rate of $\gamma = e^{-5}$. Is $f\left(\beta_1^{(t+1)}, \beta_2^{(t+1)}\right) < f\left(\beta_1^{(t)}, \beta_2^{(t)}\right)$? If not, what might be the problem?

a) $\frac{\partial L}{\partial \beta_1} = 2\beta_1 e^{\beta_1^2 - \beta_2^2 + 4\beta_2}$

$\frac{\partial L}{\partial \beta_2} = (-2\beta_2 + 4) e^{\beta_1^2 - \beta_2^2 + 4\beta_2}$

$\beta_1^{(t+1)} = 1 - 2 \cdot 0.5e^{-5} e^{1^2 - 2^2 + 4(2)} = 0$

$\beta_2^{(t+1)} = 2 - (-2(2) + 4)\, 0.5e^{-5} e^{1 - 4 + 8} = 2$

$L(1,2) = e^{1-4+8} = e^5$

$L(0,2) = e^{0-4+8} = e^4$ — loss is lower

Name: _____ ID#
(print) (last) (first)

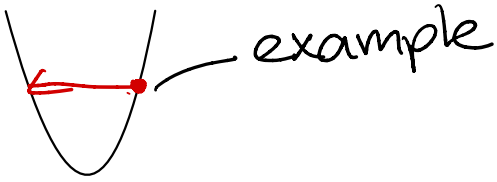
b)

$B_1^{(t+1)} = 1 - 2e^{-5} e^{1^2 - 2^2 + 4(2)} = -1$

$B_2^{(t+1)} = 2 - (2(2)+4) e^{-5} e^{1-4+8} = 2$

$\text{loss} = e^{1-4+8} = e^5$

loss is not lower → suggests learning rate is too large



Name: _____ ID#
(print) (last) (first)

4. For a given probability, p, the **odds** are $\frac{p}{1-p}$.
(a) Show that

$$p = \frac{e^{\ell}}{1+e^{\ell}} \quad \text{and} \quad 1-p = \frac{1}{1+e^{\ell}}$$

where ℓ is the **log odds**.

Consider a binary classifier that predicts the probability, p_i, for an outcome $y_i \in \{0, 1\}$.

(b) Re-write the cross entropy loss function for this classifier given a single data point / record, i.e., $\mathcal{L}_i$ from one of the previous questions, such that it is a function of the outcome, y_i, and the log odds, ℓ_i.

(c) Show that the first derivative with respect to the log odds of the function you found in the previous section equals to minus the corresponding residual.

(d) A regression model predicts $\hat{y}_i$ for outcome y_i. Show that the first derivative with respect to the prediction of a mean square error function (for this regressor) is proportional to minus the corresponding residual.

a) $1-p = 1 - \frac{e^{\ell}}{1+e^{\ell}} = \frac{1+e^{\ell}}{1+e^{\ell}} - \frac{e^{\ell}}{1+e^{\ell}} = \frac{1}{1+e^{\ell}}$

b)

$$\mathcal{L}_i \triangleq -[y_i \log p_i + (1-y_i)\log(1-p_i)]$$

$$\mathcal{L}_i = -\left(y_i \log\left(\frac{e^{\ell}}{1+e^{\ell}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\ell}}\right)\right)$$

c) $\mathcal{L}_i = -\left(y_i \log(e^{\ell}) - y_i \log(1+e^{\ell}) + 0 - (1-y_i)\log(1+e^{\ell})\right)$

$\mathcal{L}_i = -y_i \log(e^{\ell}) + y_i \log(1+e^{\ell}) + \log(1+e^{\ell}) - y_i \log(1+e^{\ell})$

Name: _____ ID#
(print) (last) (first)

$$L_i = -y_i \log(e^{\ell}) + \log(1+e^{\ell}) = -y_i \ell + \frac{1}{1+e^{\ell}}$$

$$\frac{dL}{d\ell} = -y_i + \frac{1}{1+e^{\ell}} = -y_i + p$$

d) $$MSE = \sum_{i=0}^{i} (\hat{y} - y_i)^2$$

$$\frac{dL}{d\ell} = -y_i + p = 0$$

$$y_i = p$$

$$\frac{dM}{d\hat{y}} = \sum_{i=1}^{i} 2(\hat{y} - y_i)$$

$$\downarrow = \sum_{i=1}^{i} 2(\hat{y} - p)$$