Machine Learning CS 6140
Miles Benjamin
Assignment 1

1) **Probability and Random Variables:**
1. False.
P(B ∩ A)/P(A) + P(A ∩ Bc)/P(Bc) = 1
if A and B are mutually exclusive
P(B ∩ A) = 0, P(Bc) = 1
then
P(A ∩ Bc)/P(Bc) = 1 which only holds true if P(A) = P(Bc)

2. False
If P(A) = 0
P(A|B) = 0
P(A|Bc) = 0
0+0 != 1

3. True
P(Bc u (A ∩ B) + P(Ac ∩ B) = 1
P(Bc) + P(A ∩ B) – P(Bc ∩ (A ∩ B) + P(Ac ∩ B) = 1 # P(Bc ∩ (A ∩ B) must be 0
P(Bc) + P(A ∩ B) + P(Ac ∩ B) = 1 # P(A ∩ B) + P(Ac ∩ B) = P(B)
P(Bc) + P(B) = 1 #True by definition of compliment

4. False
Since Ai is not mutually exclusive
P(Ai u Aj) = P(Ai) + P(Aj) # This is what we're trying to prove given n = 2
P(Ai) + P(Aj) – P(Ai ∩ Aj) = P(Ai) + P(Aj) # False for any time P(Ai ∩ Aj) != 0

5. True
P(A1,A2 | B1, B2) = P(A1 | B1) P(A2 | B2) = $\prod_{i=1}^{n}$ P(Ai | Bi)

2) **Discrete and Continuous Distributions:** Write down the formula of the probability density/mass functions of random variable X.
1. Multivariate Gaussian Distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

2. Laplace Distribution

$$\mathrm{Lap}(x|\mu, b) \triangleq \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$

3. Bernoulli Distribution

$$\mathrm{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

4. Multinomial Distribution

$$\mathrm{Mu}(\mathbf{x}|n, \boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

5. Dirichlet Distribution

$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K)$$

6. Uniform Distribution

$$unif(a,b) = \begin{cases} 1/(b-a) & x \in [a,b] \\ 0 & otherwise \end{cases}$$

7. Exponential Distribution

$$\mathrm{Expon}(x|\lambda) \triangleq \mathrm{Ga}(x|1, \lambda),$$

8. Poisson Distribution

$$\mathrm{Poi}(x|\lambda) = e^{-\lambda}\,\frac{\lambda^x}{x!}$$

3) **Positive-Definite Matrices:**
1. True, A is positive semidefinite
A = BtB
xtAx >= 0
xtBtBx >= 0 # Substituting in for A
(Bx)tBx >= 0 #forms two identical vectors
<Bx, Bx> >=0 # inner product of two vectors
inner product of identical vectors is always positive semi-definite

2. True. A is positive semi-definite because its eigenvalues are [8- $\sqrt{}$19, 0, 8+ $\sqrt{}$19] all of which are non-negative.

3. False. A is not positive semi-definite for the following B:

$$B = \begin{bmatrix} 1 & -100 \\ 0 & 1 \end{bmatrix} \Rightarrow A = \begin{bmatrix} 10003 & -200 \\ -200 & 3 \end{bmatrix}$$

The eigenvalues of this matrix are 10003 and -9991.

4) **Convexity of Linear Regression:**
a 1.

J(θ) = ||Y - X θ ||$_2$^2
dJ/d θ = -2Xt(Y-X θ) = -2XtY + 2XtX θ
d2J/d θ = 2XtX
2XtX is positive therefore this is convex.

a 2.

J(θ) = ||Y-X θ |||$_2$^2 + λ || θ ||$_2$^2
dJ/d θ = -2Xt(Y-X θ) + 2 λ θ
d2J/d θ = 2XtX + 2 λ
2XtX is positive and 2 λ is a constant therefore its convex

a 3.

J(θ) = ||Y-X θ ||$_2$^2 + λ || θ ||$_1$^2
if f(x) = g(x) + h(x) and both g(x) + h(x) are convex, then f(x) is convex

g(θ) = ||Y-X θ ||$_2$^2 is convex (see above)

$$h(\theta) = \lambda\|\theta\|_1 = \lambda \sum_{i=1}^{n} |\theta_i|$$

Since absolute value functions are always convex, h(θ) is convex.

Therefore J(θ) is convex

B. There can only be one X-Y pair that corresponds to the minimum, If there are multiple values of X that give the minimum value of Y then there will not be a unique solution for θ .

5) **Regression using Huber Loss:**
Batch gradient descent:

- Set δ (our learning rate) by picking a value (tune later with k-fold cross validation)
- Initialize θ to be a random vector
- While $\theta^{k+1} \mathrel{!}= \theta^k$:
  - $\theta^{k+1} = \theta^k$ - dJ/d θ | θ^k
  - dJ/d θ = -2X(½ Y - ½ θ^TX) when $\delta > |Y - \theta^T X|$
  - dJ/d θ = δX when $\delta < |Y - \theta^T X|$
- Note: $\delta = |Y - \theta^T X|$ is not continuous

Stochastic gradient descent:

- Set δ (our learning rate) by picking a value (tune later with k-fold cross validation)
- Initialize θ to be a random vector
- While $\theta^{k+1} \mathrel{!}= \theta^k$:
  - Select i = 1ϵ(X,Y)
  - $\theta^{k+1} = \theta^k$ - dJ/d θ | θ^k
  - dJ/d θ = -2x_i(½ y_i - ½ $\theta^T x_i$) when $\delta > |y_i - \theta^T x_i|$
  - dJ/d θ = δx_i when $\delta < |y_i - \theta^T x_i|$
- Note: $\delta = |y_i - \theta^T x_i|$ is not continuous

6) **PAC Confidence Bounds:**
$P(|\widehat{\theta} - \theta^0| \geq \varepsilon)$ is our confidence, therefore we can set it to 0.95 and solve.

$0.95 = 2e^{-N(0.1)^2}$

ln(0.95 / 2) = -N(0.01)

74.4 = N

Since we can't have a partial trial the answer is 75 flips.

7) **Probabilistic Regression with Prior on Parameters:**

1. Plugging the values into the normal distribution function and simplifying we get:
   $N(0,\ 1/\lambda I) = (\lambda I/\sqrt{2\pi})e^{(-\theta^2\lambda I)/2}$
2. Plugging the values into the laplace distribution function and simplifying we get:
   $Lap(\theta|0, 1/\lambda) = (\lambda/2)e^{-\lambda|\theta|}$

   The secret was remembering that the variance = 2b^2, which works out to b = 1/ λ

8) **MAP estimation for the Bernoulli with non-conjugate priors:**

1.

P(θ) = ½^(10 θ - 5) * ½^(6-10 θ)

P(D| θ) = P(D| θ =0.5)*P(θ =0.5) + P(D| θ =0.6)*P(θ =0.6)

= P(D)*P(θ =0.5) + P(D)*P(θ =0.6)

= N1/N * P(θ =0.5) + N1/N * P(θ =0.6)

I'm not 100% sure this is right, but I think it's on the right track.

2.The new prior will work better for when N is small because it takes into account that the coin might be slightly biased towards heads, neither prior will matter much when N is big because the dominant factor will be the Maximum Likelihood Estimation.

9) **Gaussian Naive Bayes:**

I don't have an answer for this question. I worked at it a long time, but in the end didn't produce anything worth showing.

10) **Linear Regression Implementation:**

See attached Jupyter Notebook file!