CS 391L Machine Learning Assignment 3

Name: Shun Zhang
Email address: jensen.zhang@utexas.edu
EID: sz4554

Problem 1

(a)

$$\begin{aligned}
Cov(y) &= E((y - E(y)(y - E(y)^T) && (1)\\
&= E((Ax + b - E(Ax + b))(Ax + b - E(Ax + b))^T) && (2)\\
&= E((Ax - E(Ax))(Ax - E(Ax))^T) && \text{Linearity of } E \quad (3)\\
&= AE((x - E(x))(x - E(x))^T)A^T && (4)\\
&= A\Sigma A^T && \text{Def. of Cov.} \quad (5)
\end{aligned}$$

(b) Base case: by the defition of eigenvalue and eigenvector, $Ax = \lambda x$.

Inductive hypothesis: assume $A^k x = \lambda^k x$ for some $k \in N$. Want to show $A^{k+1}x = \lambda^{k+1}x$.

$$\begin{aligned}
A^k x &= \lambda^k x && \text{I.H.} && (6)\\
A^{k+1} x &= A\lambda^k x && && (7)\\
A^{k+1} x &= \lambda^k Ax && && (8)\\
A^{k+1} x &= \lambda^{k+1} x && Ax = \lambda x && (9)
\end{aligned}$$

Problem 2

(a)

$$\begin{aligned}
r &= 1 - \frac{H(Y|X)}{H(X)} && (10)\\
&= \frac{H(X) - H(Y|X)}{H(X)} && (11)\\
&= \frac{I(Y|X)}{H(X)} && \text{Def. of Mutual Information} \quad (12)
\end{aligned}$$

(b) $H(Y|X) \geq 0$, $H(X) > 0$. So $\dfrac{H(Y|X)}{H(X)} \geq 0$, $1 - \dfrac{H(Y|X)}{H(X)} \leq 1$.

$H(Y|X) < H(X)$, so $\dfrac{H(Y|X)}{H(X)} <= 1$, $1 - \dfrac{H(Y|X)}{H(X)} \geq 0$.

Therefore, $0 \leq r \leq 1$.

(c) $\dfrac{H(Y|X)}{H(X)} = 1$,

Problem 3

(a) $\tanh x = \dfrac{\sinh x}{\cosh x} = \dfrac{e^x - e^{-x}}{e^x + e^{-x}} = \dfrac{1 - e^{-2x}}{1 + e^{-2x}}$. Compared to $\dfrac{1}{1 + e^{-x}}$, tanh has a steeper slope.

To be an appropriate sigmoid function, tanh needs to be scaled to range of [0, 1].

(b) We know W^k is 2×2.

$$\frac{\partial H}{\partial w_{ij}^k} = \frac{\partial}{\partial w_{ij}^k} \sum_{k=0}^{K-1} (\lambda^{k+1})^T g(W^k x^k) \tag{13}$$

$$= \frac{\partial}{\partial w_{ij}^k} (\lambda_1^{k+1} g(w_{11} x_1^k + w_{12} x_2^k) + \lambda_2^{k+1} g(w_{21} x_1^k + w_{22} x_2^k)) \tag{14}$$

So, generally,

$$\frac{\partial H}{\partial w_{ij}^k} = \lambda^{k+1} \frac{\partial}{\partial w_{ij}^k} g(W^k x^k) \tag{15}$$

$$= \lambda^{k+1} \frac{\partial}{\partial w_{ij}^k} g(\sum_l w_{il} x_l^k) \tag{16}$$

$$= \lambda^{k+1} x_j^k g'(w_i^k x^k) \tag{17}$$

Problem 4

(a) The code for computing these results is attached separately.

IG(Color) = 0.1043
IG(Size) = 0.4086
IG(Noise) = 0.0207

For small size,

IG(Color) $= 0.3219$
IG(Noise) $= 0.0207$

For medium size,

IG(Color) $= 0.1226$
IG(Noise) $= 0.1226$

For large size, IG is clearly 0.

The decision tree is

- Size
  - *Small*: Color
    - *Red*: Noise
      - *Loud*
      - *Quiet*
    - *Blue*: Noise
      - *Loud*
      - *Quiet*
  - *Medium*: Color
    - *Red*: Noise
      - *Loud*
      - *Quiet*
    - *Blue*: Noise
      - *Loud*
      - *Quiet*
  - *Medium*: **No**

(b) If the event of missing a datum is uniformly random over all the attributes, then it doesn't harm if we simply delete that line.

Problem 5

(a) $A = \pi r s + \pi r^2$. Therefore,

$H = \frac{1}{3}\pi r^2 h + \lambda(\pi r s + \pi r^2 - A)$

$\frac{\partial H}{\partial r} = \frac{2}{3}\pi r h + \lambda(\pi(s + r\frac{1}{2s}2r) + 2\pi r) = 0$

$\frac{\partial H}{\partial h} = \frac{1}{3}\pi r^2 + \lambda(\pi r\frac{1}{2s}2h) = 0$