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 $A^{k+1}x = \lambda^{k+1}x$.

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

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) $r = 0$ when two variables are independent. $r = 1$ when two vairables are perfectly correlated.

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.

$$\begin{aligned}
\frac{\partial H}{\partial w_{ij}^k} &= \frac{\partial}{\partial w_{ij}^k} \sum_{k=0}^{K-1} (\lambda^{k+1})^T g(W^k x^k) && (13)\\
&= \frac{\partial}{\partial w_{ij}^k} (\lambda_1^{k+1} g(w_{11} x_1^k + w_{12} x_2^k) + \lambda_2^{k+1} g(w_{21} x_1^k + w_{22} x_2^k)) && (14)\\
&= \lambda_1^{k+1} \frac{\partial}{\partial w_{ij}^k} g(w_{11} x_1^k + w_{12} x_2^k) + \lambda_2^{k+1} \frac{\partial}{\partial w_{ij}^k} g(w_{21} x_1^k + w_{22} x_2^k) && (15)
\end{aligned}$$

Therefore, for $w_{11}, w_{12}, w_{21}, w_{22}$, there is $\dfrac{\partial H}{\partial w_{ij}^k} = \lambda^{k+1} x_j^k g'(w_i^k x^k)$.

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,

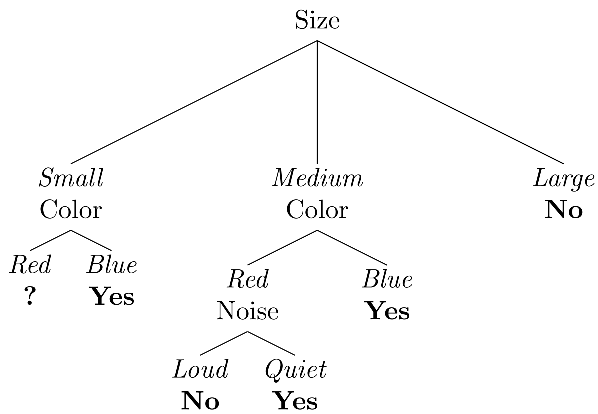
IG(Color) = 0.3219
IG(Noise) = 0.0207

For medium size,

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

For large size, IG is clearly 0.

The decision tree is



(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) $\min \dfrac{1}{3}\pi r^2 h$ such that $A = \pi r s + \pi r^2$, $s^2 = r^2 + h^2$.

Use Lagrange multiplier, $H = \frac{1}{3}\pi r^2 h + \lambda(\pi r s + \pi r^2 - A) + \gamma(r^2 + h^2 - s^2)$.

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

$\frac{\partial H}{\partial h} = \frac{1}{3}\pi r^2 + 2\gamma h = 0$