

# CS 391L Machine Learning Assignment 3

Name: Shun Zhang

Email address: `jensen.zhang@utexas.edu`

EID: sz4554

## Problem 1

(a)

$$\text{Cov}(y) = E((y - E(y))(y - E(y))^T) \quad (1)$$

$$= E((Ax + b - E(Ax + b))(Ax + b - E(Ax + b))^T) \quad (2)$$

$$= E((Ax - E(Ax))(Ax - E(Ax))^T) \quad \text{Linearity of } E \quad (3)$$

$$= AE((x - E(x))(x - E(x))^T)A^T \quad (4)$$

$$= A\Sigma A^T \quad \text{Def. of Cov.} \quad (5)$$

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

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

$$A^k x = \lambda^k x \quad \text{I.H.} \quad (6)$$

$$A^{k+1}x = A\lambda^k x \quad (7)$$

$$A^{k+1}x = \lambda^k Ax \quad (8)$$

$$A^{k+1}x = \lambda^{k+1}x \quad Ax = \lambda x \quad (9)$$

## Problem 2

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

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

(b)  $1 - \tanh^2$ .

(c) We know  $W^k$  is  $2 \times 2$ .

$$\frac{\partial H}{\partial w_{ij}^k} = \frac{\partial}{\partial w_{ij}^k} \sum_{k=0}^{K-1} (\lambda^{k+1})^T g(W^k x^k) \quad (10)$$

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

So, generally,

$$\frac{\partial H}{\partial w_{ij}^k} = \lambda^{k+1} \frac{\partial}{\partial w_{ij}^k} g(W^k x^k) \quad (12)$$

$$= \lambda^{k+1} \frac{\partial}{\partial w_{ij}^k} g\left(\sum_l w_{il} x_l^k\right) \quad (13)$$

$$= \lambda^{k+1} x_j^k g'(w_i^k x^k) \quad (14)$$

### Problem 3

(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

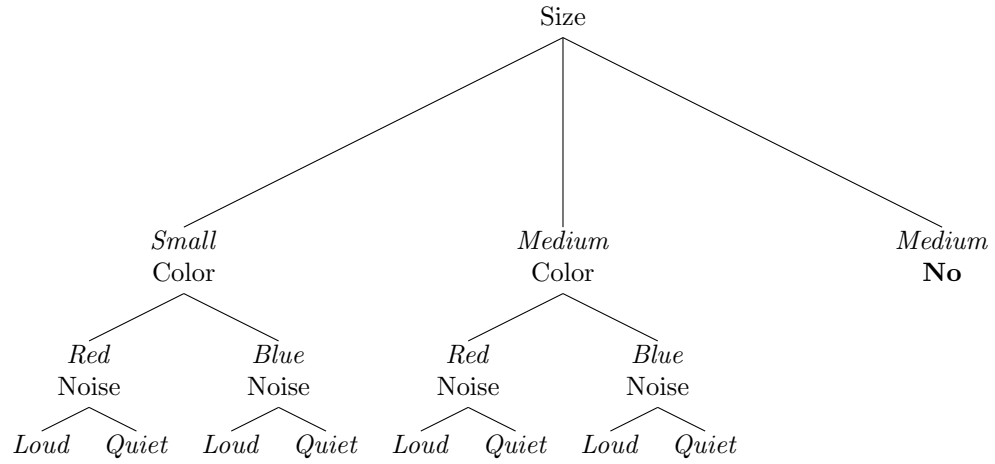
For medium size,

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 4

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

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

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

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