Kyle McGraw

CS 155 Final

1.
A) True

B) B

C) K_2, K_1, K_3

D) 2

E) B

F) False

G) B

H) False

I) True

J) False

K) B

L) True

M) True

N) True

2. $P(\text{happy}, \text{grade}, \text{year}) = P(\text{happy})\,P(\text{grade} \mid \text{happy})\,P(\text{year} \mid \text{happy})$

1)

$$P(\text{grade} = A \mid \text{happy} = \text{yes}) = \frac{1 + \sum_{(x,y)} 1_{[x_{\text{grade}} = A \wedge y = \text{yes}]}}{2 + \sum_{(x,y)} 1_{[y = \text{yes}]}}$$

happy	P(happy)
yes	0.5
no	0.5

happy	grade	P(grade|happy)
yes	A	2/3
Yes	C	1/3
no	A	1/3
no	C	2/3

happy	year	P(grade|happy)
yes	Senior	2/3
Yes	Frosh	1/3
no	Senior	1/2
no	Frosh	1/2

$$P(grade=A|happy=yes) = \frac{1+3}{2+4} = \frac{4}{6} = \frac{2}{3}$$

$$P(grade=C|happy=yes) = \frac{1+1}{2+4} = \frac{2}{6} = \frac{1}{3}$$

$$P(grade=A|happy=no) = \frac{1+1}{2+4} = \frac{2}{6} = \frac{1}{3}$$

$$P(grade=C|happy=no) = \frac{1+3}{2+4} = \frac{4}{6} = \frac{2}{3}$$

$$P(year=senior|happy=yes) = \frac{1+3}{2+4} = \frac{4}{6} = \frac{2}{3}$$

$$P(year=frosh|happy=yes) = \frac{1+1}{2+4} = \frac{2}{6} = \frac{1}{3}$$

$$P(year=senior|happy=no) = \frac{1+2}{2+4} = \frac{3}{6} = \frac{1}{2}$$

$$P(year=frosh|happy=no) = \frac{1+2}{2+4} = \frac{3}{6} = \frac{1}{2}$$

2)

$$P(year=frosh, grade=C, happy=no)$$

$$= P(happy=no)\, P(grade=C|happy=no)\, P(year=frosh|happy=no)$$

$$= 0.5 \cdot 2/3 \cdot 1/2$$

$$= 1/6$$

3)

```
R = random()
if R < P(happy)
   h = no
else
   h = yes
R = random()
if R < P(grade=C | happy=h)
   g = C
else
   g = A
R = random()
if R < P(year=frosh | happy=h)
   y = frosh
else
   y = Senior
return {happy: h, grade: g, year: y}
```

3.

$$\tilde{x} = Ax$$

1)

$$w^T x = \tilde{w}^T \tilde{x}$$
$$w^T x = \tilde{w}^T A x$$
$$w^T = \tilde{w} A$$
$$w = A^T \tilde{w}$$

2)

$$\text{argmin}_{\tilde{w}} \frac{\lambda}{2} \|\tilde{w}\|^2 + \sum_i (y_i - \tilde{w}^T \tilde{x}_i)^2$$

$$\tilde{w} = A^{T^{-1}} w$$

$$\text{argmin}_{A^{T^{-1}}w} \frac{\lambda}{2} \|A^{T^{-1}} w\|^2 + \sum_i (y_i - w^T x_i)^2$$

$$\text{argmin}_w = \frac{\lambda}{2} \|A^{T^{-1}} w\|^2 + \sum_i (y_i - w^T x_i)^2$$

3)

This answer is different from standard ridge regression because of the scaling matrix in the regularization term. Since its the inverse, we will have the weights in the regularization inversly scaled as we minimize them.

4.

$$P(s'|s) = \frac{e^{-\|u(s') - v(s)\|_2^2}}{Z(s)}$$

u, v: song $\rightarrow$ vector maps

Z: normalization

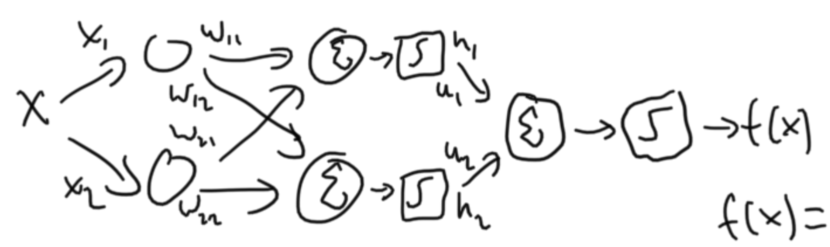
$$P(s'|s) = \frac{e^{-\|x(s') - x(s)\|_2^2}}{Z(s)}$$

1)

We know that $P(L)$ for the dual-point model will never be less than that for the single-point model because there are two vectors, U and V, representing each song as opposed to one, X. Therefore, if we choose U and V to maximize $P(L)$, we know we can always set both to X to get the same $P(L)$ as the single-point model, but could also have choices for U and V that would have a larger $P(L)$.

2) If these equations are equal than we have $U = V = X$.

5.



$$f(x) = \sigma\left(\sum_{i=1}^{2} u_i \, \sigma\left(\sum_{j=1}^{2} w_{ji} x_j\right)\right)$$

1. $\frac{\partial}{\partial w_{11}} L(y, f(x)) = \frac{\partial}{\partial w_{11}} (y - f(x))^2$

$$= \frac{\partial (y-f(x))^2}{\partial f(x)} \frac{\partial \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)}{\partial \sum_{i=1}^{2} u_i h_i(x)} \frac{\partial \sum_{i=1}^{2} u_i h_i(x)}{\partial h_1(x)} \frac{\partial \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)}{\partial \sum_{j=1}^{2} w_{j1} x_j} \frac{\partial \sum_{j=1}^{2} w_{j1} x_j}{\partial w_{11}}$$

$$= -2(y - f(x)) \cdot \left(\sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right) \cdot \left(1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right)\right) \cdot u_1$$
$$\cdot \left(\sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right) \cdot \left(1 - \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\right)\right) \cdot x_1$$

2. $x_1 = 0.1 \quad x_2 = 0.5 \quad y = 0.75 \quad u_1 = 0.5 \quad u_2 = -0.1$

$w_{11} = 0.25 \quad w_{12} = 0.1 \quad w_{21} = 0.05 \quad w_{22} = -0.25$

$$\sum_{j=1}^{2} w_{j1} x_j = 0.25 \cdot 0.1 + 0.05 \cdot 0.5 = 0.05$$

$$\sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right) = 0.512$$

$$\sum_{j=1}^{2} w_{j2} x_j = 0.1 \cdot 0.1 - 0.25 \cdot 0.5 = -0.115$$

$$\sigma\left(\sum_{j=1}^{2} w_{j2} x_j\right) = 0.471$$

$$\sum_{i=1}^{2} u_i h_i(x) = 0.05 \cdot 0.512 - 0.115 \cdot 0.471 = 0.209$$

$$\sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right) = 0.552$$

$$-2(y - f(x)) \cdot \left(\sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right) \cdot \left(1 - \sigma\left(\sum_{i=1}^{2} u_i h_i(x)\right)\right)\right) \cdot u_1$$
$$\cdot \left(\sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right) \cdot \left(1 - \sigma\left(\sum_{j=1}^{2} w_{j1} x_j\right)\right)\right) \cdot x_1 = -0.00122$$

3. The vanishing gradient problem arises from the derivative of $f(x)$. Since the weights of all the previous layers are involved in calculating $f(x)$ and the sigmoid function values are between 0 and 1, we get a product of a lot of small numbers. As we increase the amount of layers, we increase the number of small numbers multiplied together, decreasing the gradient.