(1)

(A) $\boxed{\text{True}}$ → some λ exists where L1-regression regularized has same solution as given constrained Regression

(B) $\boxed{D}$ → $(1-\frac{1}{N})$ → $P(\text{No } x,y)$ → N samples per bootstrap, M bootstraps $M \cdot N$

(C) $\boxed{K1}$, K3, K2

(D) $\boxed{\hat{y} = 4}$

(E) $\boxed{A}$ ⇒ same application of weight 4 in (2) encourages correlation

(F) $\boxed{\text{False}}$ ⇒ Hard margin SVM just won't converge not $w=0$

(G) $\boxed{A}$ ⇒ support vectors remain the same, so no change in w, b

(H) $\boxed{\text{False}}$ ⇒ new predictions can change to incorrect or not change previous predictions

(I) $\boxed{\text{True}}$ ⇒ Linear algebra operations provide closed form solution with 2 matrices fixed

(J) $\boxed{\text{False}}$ ⇒ Bias variance tradeoff shows test error not minimized with var or bias min.

(K) $\boxed{B}$ ⇒ $\prod_{j=2}^{M} P(x^j | y^j) P(y^j | y^{j-1})$ same for all values, each model learns equal

(L) $\boxed{\text{True}}$ ⇒ Exists a tradeoff between accuracy and size/variability of training space

(M) $\boxed{\text{True}}$ ⇒ With infinite data any classifier function can be learned with decision trees

(N) $\boxed{\text{True}}$ ⇒ Infinite data means ~~mod~~ infinitely harder for model to fit on specific points and not ~~generalize~~ generalize

NAIVE BAYES

(1) $P(\text{Grade} = A \mid \text{Happy} = \text{Yes}) = 2/3$ $\quad P(\text{Grade} = A \mid \text{Happy} = \text{No}) = 1/3$

$P(\text{Grade} = C \mid \text{Happy} = \text{Yes}) = 1/3$ $\quad P(\text{Grade} = C \mid \text{Happy} = \text{No}) = 2/3$

$P(\text{Year} = \text{Freshman} \mid \text{Happy} = \text{Yes}) = 1/3$ $\quad P(\text{Year} = \text{Freshman} \mid \text{Happy} = \text{No}) = 1/2$

$P(\text{Year} = \text{Senior} \mid \text{Happy} = \text{Yes}) = 2/3$ $\quad P(\text{Year} = \text{Senior} \mid \text{Happy} = \text{No}) = 1/2$

$P(\text{Happy} = \text{Yes}) = 1/2$ $\quad P(\text{Happy} = \text{No}) = 1/2$

(2) $P(Y, G, H) = P(H) \cdot P(Y \mid H) \cdot P(G \mid H)$

$= P(\text{Happy} = \text{No}) \cdot P(\text{Freshman} \mid \text{Not happy}) \cdot P(C \mid \text{Not happy})$

$= 1/2 \cdot 1/2 \cdot 2/3 = 2/12 = 1/6$

$$\boxed{P(\text{Year} = \text{Freshman}, \text{Grade} = C, \text{Happy?} = \text{No}) = 1/6}$$

(3)

```
if random() > P(Happy? = No)
[redacted]
    happy = True
else:
    happy = False
if random() > P(Grade = C | Happy? = happy)
    grade = A
else:
    grade = C
if random() > P(Year = Freshman | Happy? = happy)
  year ~~grade~~ = Senior
else:
  year ~~grade~~ = Freshman

return happy, grade, ~~Fresh~~ year
```

Random() represents generating probabilities via random number [0, 1]

True = Yes, False = No

Probabilities can be called from training results

(3) DATA TRANSFORMATIONS

(1) $w^T x = \hat{w}^T \hat{x}$, given $\hat{x} = Ax$ we can substitute

$w^T x = \hat{w}^T A x$ which implies $w^T = \hat{w}^T A$

$\Rightarrow (w^T)^T = (\hat{w}^T A)^T \Rightarrow w = A^T \hat{w}$

so we get w as a function of A and $\hat{w}$ where:

$$\boxed{w = A^T \hat{w}}$$

(2) given $w = A^T \hat{w} \Rightarrow \hat{w} = (A^T)^{-1} w \Rightarrow \hat{w}^T = w^T A^{-1}$

$$\Rightarrow \arg\min_{\hat{w}} \frac{\lambda}{2} \|\hat{w}\|^2 + \sum_i (y_i - \hat{w}^T \hat{x}_i)^2$$

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

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

optimization problem: $\arg\min_{\hat{w}} \frac{\lambda}{2} \|(A^T)^{-1} w\|^2 + \sum_i (y_i - w^T x_i)^2$

(3) Notice from (2) the squared loss ~~expression~~ term simplifies to the standard ridge regression squared loss term. Thus, the only difference is in the regularization term where $(A^T)^{-1} w$ is L2 regularized instead of just w. In context of A as a transformation on x, this essentially means the optimization from (2) changes from (1) in that the ~~sc~~ inverted scaling of the weights is regularized compared to the standard ridge.

(4)

(1) Notice that U and V can always converge to X where the dual point model now has $u, v = x$ in its optimization. Since this optimization is always a choice we can argue that the dual point model in its worst case, can always do as well as the best performance of the single-point model since the dual point can always replicate the single point. As a result, given optimal u, v, x we ~~know~~ can conclude the dual-point ~~[illegible]~~ model likelihood can only be as worse as single-point model, so the dual-point $P(S)$ likelihood is never less than the single point model.

(2) If the transition probabilities of both single-point (18) and dual point (6) models are the same it implies that $U = V$ since $u(s') - v(s') = x(s') - x(s)$ for all s.

(5)

(A) Chain rule: $\frac{\partial}{\partial w_{11}} L(y, f(x)) \equiv \frac{\partial (y-f(x))^2}{\partial w_{11}}$

$$= \frac{\partial (y-f(x))^2}{\partial f(x)} \cdot \frac{\partial f(x)}{\partial(\sum u_i h_i(x))} \cdot \frac{\partial(\sum u_i h_i(x))}{\partial h_1(x)} \cdot \frac{\partial h_1(x)}{\partial \sum w_{ji} x_j} \cdot \frac{\partial \sum w_{ji} x_j}{\partial w_{11}}$$

$$= \boxed{[-2(y-f(x))]\underbrace{[f(x)(1-f(x))]}_{\sigma(s)\text{ derivative}}[u_1]\underbrace{[h_1(x)(1-h_1(x))]}_{\sigma(s)\text{ derivative}}[x_1]}$$

(B) $h_1(x) = \sigma(w_{11}x_1 + w_{21}x_2) = \sigma(.05) = .5124$

$h_2(x) = \sigma(w_{22}x_1 + w_{22}x_2) = \sigma(-.115) = .4713$

$f(x) = \sigma(u_1 h_1(x) + u_2 h_2(x)) = \sigma((.5124)(.5) - (.4713)(.1)) = \sigma(.209) = .5521$

$\Rightarrow -2(.75-.5521)(.5521(1-.5521))(.5)(.5124(1-.5124))(.1)$

$\Rightarrow = -.001222 \Rightarrow \boxed{\frac{\partial}{\partial w_{11}} L(y, f(x)) = -.001222}$ ↳ done

(C) The sigmoid term as well as its derivatives (which include $\to \frac{\partial}{\partial \sigma(s)}\sigma(s) = \sigma(s)(1-\sigma(s))$ sigmoid terms) result in the vanishing gradient problem, notice in (A) there are $f(x)$ and $h_1(x)$ terms which are both non-linearities, non-linearities always result in output < 1 which when multiplied in chain rule will go to zero resulting in a vanishing gradient. For larger layer networks the chain rule is larger ~~resulting in more multiplications and likelihood of vanishing,~~ exacerbating the problem