1

(a) Unsupervised Learning, Clustering
You need to identify the topics by grouping the similar contents from a large collection of blogs. Features might be frequency of each word.

(b) Supervised learning, Regression
You need to give the credit limit (numerical value) based on some features, such as individual salary, age, number of bank accounts. The label would be the actual credit limit.

(c) Supervised Learning, Classification
The result is categorical (either pass or fail) and the training could be based on actual pass/fail data. Features could be midterm grade, homework grade, final exam grade. The label is 0 for fail and 1 for pass in the training data set.

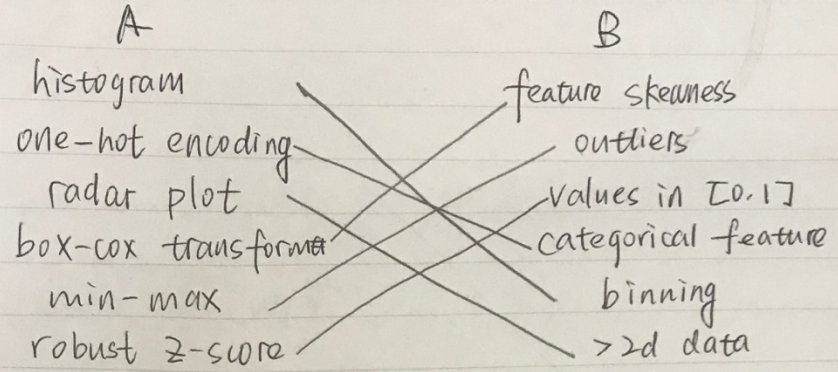
(d) Unsupervised Learning, Clustering
We don't have a label here. All we want to do is try to segment the customers. Features could be frequency of going to grocery store, purchase amount.

(e) Unsupervised Learning, Density Estimation
The probability could be given by density estimation. Features could be span email title, length of the email.

Date

2.

a.



b.

(i) nominal : Eye color (Blue, Green, Brown, Hazel, Black).

(ii) Ordinal : Rank in the competition (1, 2, 3).

(iii) numerical: Temperature (65°F, 74°F, 83°F)

	(1) frequency distribution	(2) median	(3) mean
(i) nominal	X		
(ii) ordinal		X	
(iii) numerical			X

c.

Student		From Yale	From Stanford	From CMU	From UCB	From MIT	Major Policy	Major Engineering
[Yale	Policy	1	0	0	0	0	1	0
Stanford	Policy	0	1	0	0	0	1	0
Stanford	Engineering	0	1	0	0	0	0	1
CMU	Policy	0	0	1	0	0	1	0
MIT	Engineering	0	0	0	0	1	0	1
CMU	Engineering]	0	0	1	0	0	0	1

The data now contains [7] features.

KOKUYO

No.

Date

3.

a. i. D

ii. $L(w) = \sum_{i=1}^{n} [y_i - (w_1 x_i + w_0)]^2$

$$\frac{\partial L}{\partial w_0} = -2\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i) = 0$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$\frac{\partial L}{\partial w_1} = -2\sum_{i=1}^{n}(y_i - w_0 - w_1 x_i) x_i = 0$$

$$\sum_{i=1}^{n} y_i x_i - (\bar{y} - w_1 \bar{x}) x_i - w_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y} x_i + w_1 \bar{x} x_i - w_1 x_i^2 = 0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i - \bar{y} + w_1 \bar{x} - w_1 x_i) x_i = 0$$

$$\sum_{i=1}^{n} y_i - \bar{y} + w_1(\bar{x} - x_i) = 0 \Rightarrow \sum_{i=1}^{n}(y_i - \bar{y}) = -w_1 \sum_{i=1}^{n}(\bar{x} - x_i)$$

$$w_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = (X^T X)^{-1} X^T Y$$

KOKUYO

No.

Date

b.

i. I shouldn't employ linear regression without regularization. Since we have 16,000 features, the model could be too complex if without regularization. In addition, some features with large values may dominate the model while not having huge impact on the sales in reality. It tends to overfit as we have so many features, thus harder to generalize.

ii. I could use Lasso regularization on this linear regression model. Lasso regression could be used to select features as w_i could reach zero. After regularization, we get the "critical" features and the shrinkage penalty λ constrains model complexity.

C (iii)

d. Denote y to be revenue, x_1 to be utility

$$y = w_0 + w_1 x_1 + w_2 x_1^2$$

Here w_0, w_1, w_2 are features for this polynomial model. We can still use linear regression to predict y, and the difference is x_is are no longer linear.

No.

Date

e.

i. Multi-collinearity will occur on this data. If we include all these correlated features, the wis are no longer the effect of the corresponding feature. We can not know whether feature 1 is influencing the model or it is just because the correlation of feature 2

ii. We could test the collinearity by evaluating variance inflation factor. Run regression of x_j from all other regressors. Large VIF implies that two features are hugely correlated.

f. Yes, we could use non-linear basis and generalized additive Models.

Denote y as revenue, X_1 as price, X_2 as brand perception X_3 as region.

$$y = w_0 + f_1(X_1) + f_2(X_2) + f_3(X_3)$$

↑ polynomial ↑ linear ↑ quadratic

Each function of x_i could be non-linear and we predict them using different functions, then add them together to predict revenue.

No.

Date

4. a.

	Bias	Variance
Linear Regression	High	Low
Polynomial regression w/ degree 3	Low	Low
Polynomial regression w/ degree 10	Low	High

b.

Approximation Error	Estimation Error	Computational time
Larger	Smaller	Higher

With model complexity fixed, we are changing the size of training data. With larger K, training data size becomes larger. The computational time is higher as we need to iterate the validation process for k times. The model tends to overfit on small data set. So for larger k bias will be larger (approx. error) and variance (est. error) will be smaller.

KOKUYO