Part 1: Multiple choice questions

There may be one or two correct answer(s) for each question. Choose all correct answers for each question.

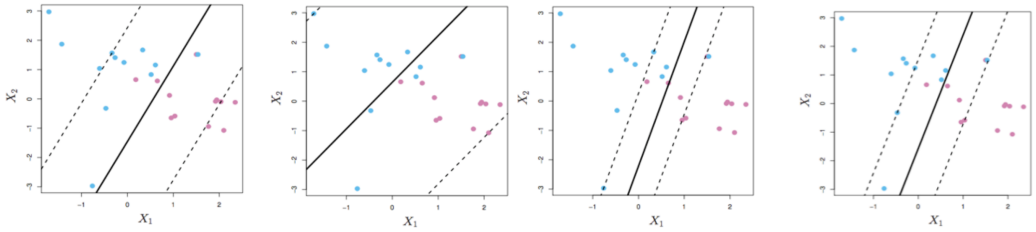
1. Which of the following support vector machine decision boundaries and margins correspond to the smallest *C* value of:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

a. b. c. d.



2. Which of the following algorithms are deterministic algorithms? 输入一定，输出也一定
   a. Multiple regression
   b. PCA
   c. t-SNE
   d. k-means 初始化中心点不同，结果也不同
3. Which of the following are considered ensemble methods?
   a. Bagging and boosting trees
   b. K-fold cross validation and penalized goodness of fit.
   c. Lasso and Ridge Regression
   d. Maximum Likelihood estimation and nonparametric kernel density estimation.

Part 2: Sample short answer question

Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergraduate weighted average mark, X_3 = years of previous programming experience and Y = receive a High Distinction (HD). We decide to use logistic regression to solve this classification problem. Below is the R code and results:

```
> lr.results <- glm(Grade ~ ., data = students, family = binomial )

> summary(lr.results)

Call:
```

```
glm(formula = Grade ~ ., family = binomial, data = students)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.08298  -0.23113  -0.10685  -0.01107   2.66679

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -17.72849    8.02049  -2.210   0.0271 *
UgradMark    -0.03386    0.12853  -0.263   0.7922
HrsStudied    0.72599    0.14819   4.899 9.63e-07 ***
YearsProg     0.52373    0.43569   1.202   0.2293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 140.45  on 129  degrees of freedom
Residual deviance:  42.97  on 126  degrees of freedom
AIC: 50.97

Number of Fisher Scoring iterations: 7
```

a. Write down the equation of the logistic regression model (include the coefficients from the R outputs).
b. Which predictor variable best explains the grade response variable? Justify your answer briefly.
c. For a student with undergraduate mark of 65, 20 hours of studying, and no programming experience, what is the probability the student will obtain a HD result?
d. Briefly describe how to interpret the coefficients of the logistic regression.
e. We want to extend the problem to classify students into all grades – Fail, Pass, Credit, Distinction and High Distinction. Is logistic regression a suitable algorithm for this problem? If not, suggest another more suitable algorithm.

Part 3: Long answer question

Describe how you would solve the following problem using Monte Carlo simulation. You may use pseudo code as part of your answer.

“The coupon collectors problem”
Every time you go to a supermarket and spend over $30 you get given a collectable item. There are 20 items to collect. How many $30 shops do you need to do to collect all 20 collectable items with probability of over 95%?