Part 1: Multiple choice questions

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

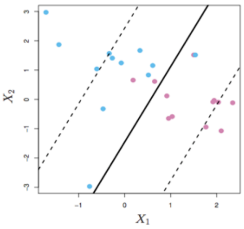
1. Which of the following support vector machine decision boundaries and margins correspond to the smallest *C* value of:

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

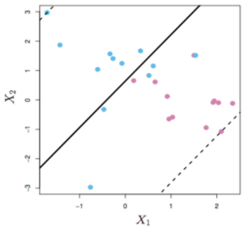
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

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

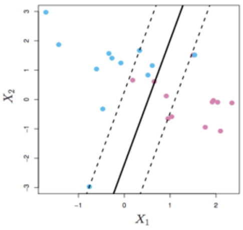
a.



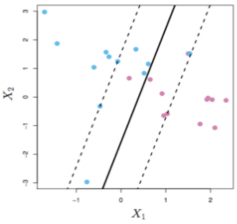
b.



[✓ c.]



d.



2. Which of the following algorithms are deterministic algorithms?
   a. [✓ Multiple regression]
   b. [✓ PCA]
   c. t-SNE
   d. k-means
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).
p = probability of positive class (receiving a HD)
*log(p/(1 - p)) = -17.72 - 0.034 (UgradMark) + 0.72 * HrsStudied + 0.52 YearsProg*
where
- *UgradMark = students undergrad mark*
- *HrsStudied = hours studied the subject*
- *YearsProg = Years of programming experience*

R^2, p-value

b. Which predictor variable best explains the grade response variable? Justify your answer briefly.
- *HrsStuded best explains the chance of getting a HD as it is the only statistically significant predictor with the other predictors having large p-value (no statistical significance).*

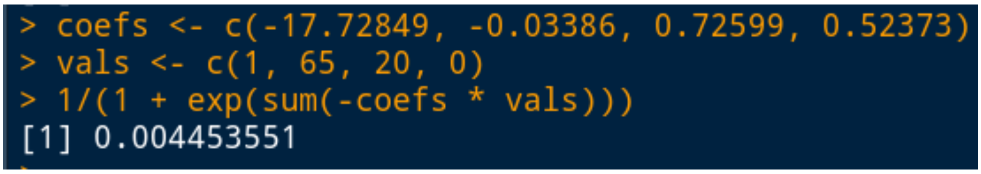
c. For a student with undergraduate mark of 65, 20 hours of studying, and no programming experience, what is the probability the student will obtain a HD result?
This is a simple calculation problem using the equation above or the direct equation
p = 1/(1 + exp(-Xbeta)) = 0.004453551 (see r screenshot below)

$-17.72 - 0.034 \times 65 + 0.72 \times 20 + 0$

$= \cancel{5.53}$

$= -5.4959$

$p = \frac{1}{1+e^{5.4959}} = 4.4535 \times 10^{-3}$



```
> coefs <- c(-17.72849, -0.03386, 0.72599, 0.52373)
> vals <- c(1, 65, 20, 0)
> 1/(1 + exp(sum(-coefs * vals)))
[1] 0.004453551
```

d. Briefly describe how to interpret the coefficients of the logistic regression.

From the equation above, it represents the increase in the log odds (log (p/(1-p)) for each unit increase in the predictor. Or looking at the odds ratio.

*p/(1-p) = exp(X beta) = exp(-17.72 - 0.034 (UgradMark) + 0.72 * HrsStudied + 0.52 YearsProg). So each coefficient represents the increase or decrease in odds.*

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.

Logistic regression is a binary classifier and cannot handle multiclass data. Another algorithm is more suitable, such as multinomial logistic regression or fitting many submodels such as one vs all dicussed in lectures (give detail if you were answering this)

Part 3: Long answer question

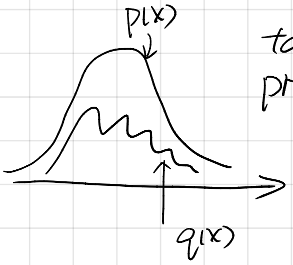
Describe how you would solve the following problem using Monte Carlo simulation. You may use pseudo code as part of your answer.

“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%?

This problem was partially solved in the lecture where the Monte Carlo simulation was used to generate the distribution of the number of shops required to collect all Coles items. Here there the extension is to determine the value on the distribution to be at the 95th quantile. So very briefly, an algorithm needs to be setup where a set of 20 items is defined (could be integers, letters, doesnt matter). A counter is setup to draw a value from the set of 20 which represents a $30 shopping trip at the super market and an item gained from the purchase. The counter iterates and a new value is drawn from the item set until the set of drawn items includes all 20 items. Once all 20 items have been drawn, the counter is retained and recorded and represents the number of shops required for one successful collection of all items. This process is repeated m times for some large m to represent m different realisations of the scenario and giving m different number of shopping trips required. These m values can be used to estimate the distribution of the number of shopping trips required. This can be done with a probability history or a density estimate. To determine the number of trips required to be confident of a 95% chance of having all the collector items, one would determine the value under the histogram or density such that the area to the left of that value is 95% or 0.95.

rejection sampling

target: $p(x)$
proposal: $q(x)$



```
i = 1
while $i \neq N$ =
        $x^{(i)} \sim q(x)$      proposal distribution
        $u \sim U(\underset{\text{mean}}{\underset{\uparrow}{0}}, \underset{\text{std}}{\underset{\uparrow}{1}})$
        if $u <$ Target distribution $\frac{p(x^{(i)})}{mq(x^{(i)})}$ then allept $x^{(i)}$
        i += 1
        else reject $x^{(i)}$
        end if
end while
```

Metro

```
for i = 1 to N
    $u \sim Nor(0,1)$
    $x^* \simeq q(x^* | x^{(i)})$
    if $u < \min\{1, \frac{p(x^*)q(x^{(i)} | x^*)}{p(x^{(i)})q(x^* | x^{(i)})}$
    $x^{i+1} = x^*$
    else
        $x^{i+1} = x^{(i)}$
```

PCA Kmeans T-SNE