1. Bias-Variance Trade-off. Multiple choices. Each question may have more than one correct answer.

   (a) (3 points) In general, which of the following regression models is expected to achieve the smallest variance for prediction if the model provides a consistent estimation of the truth?

   A. Linear regression
   B. Nadaraya–Watson estimator
   C. Smoothing spline
   D. k nearest neighbor

   (b) (3 points) Which of the following statement(s) is true for 1-nearest neighbor?

   A. It can achieve small variance
   B. It can achieve small bias
   C. At a target point x_0 where $P(Y = 1|X = x_0)$ is almost 1, 1-nearest neighbor classification makes low prediction error
   D. 1-nearest neighbor regression is similar to kernel estimator with small bandwidth

   (c) (3 points) Which of the following parameters can be used to tune the bias-variance trade-off

   A. λ in Lasso and Ridge
   B. k in nearest neighbor
   C. number of folds k in cross-validation
   D. α in regularized QDA for estimating the covariance matrix $\widehat{\Sigma}_k(\alpha) = \alpha\widehat{\Sigma}_k + (1-\alpha)\widehat{\Sigma}$ (or, see HW5 Q2b)
   E. number of trees T in boosting

   (d) (3 points) More bias is likely to happen at which of the following scenario?

   A. Predicting at a boundary point instead of an interior point when the Nadaraya–Watson estimator is used
   B. Knots positions are not optimized when using a quadratic spline
   C. Using EM algorithm instead of gradient descent when optimizing a likelihood function
   D. Using logistic regression instead of k nearest neighbor

   (e) (3 points) What are the consequences if both bias and variance of a model converge to 0

   A. The model has optimal training error
   B. The model has optimal prediction error on independent testing data
   C. The model is consistent
   D. The model could be inconsistent

2. Unsupervised Learning. Multiple choices. Each question may have more than one correct answer.

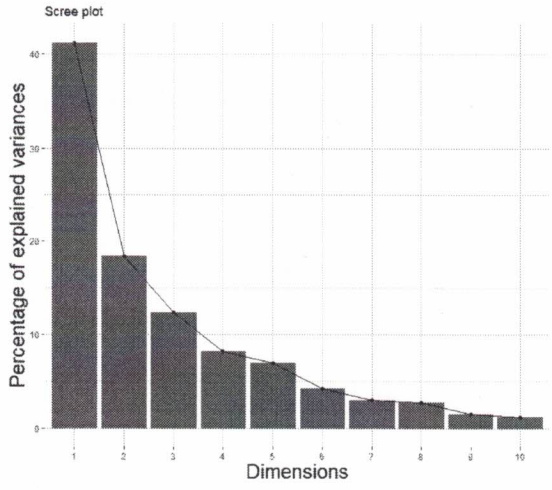
   (a) (3 points) Which of the following scenarios can benefit from performing PCA analyses?

   A. Reducing the dimensionality of the data
   B. Visualizing the data in a low dimensional plot
   C. Finding the most significant variable that associated with the outcome
   D. Detecting outliers
   E. Finding the variable with the largest marginal variance

   (b) (3 points) Which of the following method can guarantee to find the best clustering result that minimizes the within-cluster variation?

   A. Combinatorial algorithm
   B. k means
   C. Hierarchical clustering
   D. Multidimensional scaling
   E. Self-organizing maps

   (c) (3 points) A researcher collected a dataset with 200 observations, 1000 covariates (X) and a continuous outcome Y. PCA was performed on X and the following plot was obtained. Which of the following statements is true?



   A. The data lies approximately in a low dimensional sub-space
   B. A nearest neighbor regression will perform well on this data
   C. If Y depends only on the first several principal components, a ridge regression will perform well on this data
   D. If Y depends only on the first several principal components, a Lasso regression will perform well on this data
   E. If Y depends only on the first several principal components, a nearest neighbor regression will perform well on this data

(d) (3 points) Which of the following algorithms was covered in our lecture

A. Self-organizing map
B. UMAP
C. Spectral Clustering
D. tSNE
E. Principal Coordinates Analysis

3. Suppose we fit a linear (or penalized linear) regression using the observed data: $\mathbf{X}_{n\times p}$ and $\mathbf{y}_{n\times 1}$, with $n > p$. Suppose the true underlying model is given by $Y = X\boldsymbol{\beta} + \epsilon$ with ϵ's follows iid normal distribution, and $\mathbf{X}$ has full column rank. Let $\widehat{\boldsymbol{\beta}}^{\text{OLS}}$, $\widehat{\boldsymbol{\beta}}^{\text{Lasso}}$ and $\widehat{\boldsymbol{\beta}}^{\text{Ridge}}$ denote the OLS, lasso and ridge estimators, respectively on this same data. For ridge and Lasso, we assume that λ is a nonzero constant. Answer the following questions.

(a) (3 points) Select the smallest and the largest items from the list below ($\|\cdot\|$ denotes ℓ_2 norm)

A. $\|y - X\widehat{\boldsymbol{\beta}}^{\text{OLS}}\|^2$
B. $\|y - X\widehat{\boldsymbol{\beta}}^{\text{Lasso}}\|^2$
C. $\|y - X\widehat{\boldsymbol{\beta}}^{\text{Ridge}}\|^2$
D. $\|y\|^2$

(b) (3 points) What is true regarding the Hat-matrix?

A. The rank is p
B. The rank is n
C. $\mathbf{H}(y - X\widehat{\boldsymbol{\beta}}^{\text{OLS}}) = \mathbf{0}$
D. $\mathbf{H}(y - X\widehat{\boldsymbol{\beta}}^{\text{Lasso}}) = \mathbf{0}$
E. $\mathbf{H}(y - X\widehat{\boldsymbol{\beta}}^{\text{Ridge}}) = \mathbf{0}$

(c) (3 points) Suppose we run a ridge regression with just one column of X, e.g., regress $\mathbf{y}$ on X_1 and get coefficient β_1. We now include an exact copy $\tilde{X}_1 = X_1$, and regress $\mathbf{y}$ on both $\tilde{X}_1$ and X_1. We would expect the two coefficients to be

A. One of them is the same as β_1, the other one is 0
B. The two are exactly the same
C. Ridge regression cannot obtain a proper solution in this case
D. They will be the same as the Lasso regression solution with some other λ value
E. Any situation could happen, depends on the correlation between $\mathbf{y}$ and X_1

(d) (3 points) The Lasso regression solution path is the same as

A. Forward stepwise regression
B. Forward stagewise regression
C. Backward stepwise regression
D. LARS
E. Best subset selection

$\lambda \to 0 \Rightarrow \hat{\beta}^{OLS}$

(e) (3 points) When we decrease the turning parameter λ of Lasso from infinity to 0, the residual sum of squares of the training data will

A. Steadily increase
B. Steadily decrease
C. Remain constant
D. Increase initially, and then start decreasing
E. Decrease initially, and then start increasing

4. Other topics. Multiple choices. Each question may have more than one correct answer.

(a) (3 points) Select the models with the smallest and the largest degrees of freedom.

A. Cubic spline with 4 knots 8
B. Quadratic spline with 6 knots 9
C. Linear spline with 8 knots 10
D. Natural cubic spline with 11 knots 11
E. Piecewise linear with 5 knots 12

(b) (3 points) Consider two curves $\hat{g}_1$ and $\hat{g}_2$, defined by

$$\hat{g}_1 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(3)}(x) \right]^2 dx \right)$$

$$\hat{g}_2 = \arg\min_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(4)}(x) \right]^2 dx \right)$$

where $g^{(m)}$ represents the mth derivative of g.

A. As $\lambda \to \infty$, $\hat{g}_1$ have the smaller training RSS
B. As $\lambda \to \infty$, $\hat{g}_2$ have the smaller training RSS
C. As $\lambda \to 0$, $\hat{g}_1$ have the smaller testing RSS
D. As $\lambda \to 0$, $\hat{g}_2$ have the smaller testing RSS

$\hat{g}_2$ is more flexible

(c) (3 points) Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First, we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next, we use 1-nearest neighbors (i.e. $k = 1$) and get an average error rate (averaged over both training and testing data sets) of 18%. Based on these results, which method should we prefer to use for the classification of new observations?

A. Logistic regression
B. 1-nearest neighbor
C. A combination of the two
D. Cannot decide

1NN gives 0 training error.
So testing error is 36%

(d) (3 points) Increasing λ in Lasso will

A. Increase the bias
B. Decrease the bias
C. Increase the variance
D. Decrease the variance

(e) (3 points) In a genetic study, with possibly millions of SNPs being measured, which of the following models may greatly suffer in terms of its performance

A. Best subset selection
B. Lasso
C. Ridge
D. Fused Lasso
E. Nadaraya–Watson kernel estimator

(f) (3 points) Which of the following is computationally the easiest (when both n and p are large)? Choose only one answer.

A. Best subset selection
B. Lasso
C. Nonliear SVM
D. Boosting with deep trees
E. Deep learning

EM.

$$\ell(\theta; X) = N_{AA}\log(P_A^2) + N_{BB}\log(P_B^2) + N_O\log(P_O^2) + N_{AB}\log(2P_AP_B) + N_{AO}\log(2P_AP_O) + N_{BO}\log(2P_BP_O) + \log\left(\frac{n!}{N_{AA}!\,N_{AO}!\,N_{AB}!\,N_{BB}!\,N_{BO}!\,N_O!}\right).$$

Given N_A, N_{AO} and N_{AA} are two possible outcomes out of N_A trails, and their prob are

$$\frac{P_A^2}{P_A^2+2P_AP_O} \quad \text{and} \quad \frac{2P_AP_O}{P_A^2+2P_AP_O} \quad \text{respectively.}$$

Hence

$$N_{AA} \mid N_A \sim \text{Binomial}\left(N_A, \frac{P_A^2}{P_A^2+2P_AP_O}\right)$$

M step. since $P_O = 1 - P_A - P_B$

$$\text{set } 0 = \frac{\partial L}{\partial P_A} = N_{AA} \cdot 2 \cdot \frac{1}{P_A} + N_O \cdot 2 \cdot \frac{-1}{1-P_A-P_B} + N_{AB} \cdot \frac{1}{P_A} + N_{AO} \cdot \frac{1}{P_A} + N_{BO} \cdot \frac{-1}{1-P_A-P_B}$$

$$N_{AO} \frac{1-2P_A-P_B}{P_A(1-P_A-P_B)} + N_{BO} \frac{-1}{1-P_A-P_B}$$

$$\text{set } 0 = \frac{\partial L}{\partial P_B} = N_{BB} \cdot 2 \cdot \frac{1}{P_B} + N_O \cdot 2 \cdot \frac{-1}{1-P_A-P_B} + N_{AB} \frac{1}{P_B} + N_{BO} \frac{1-2P_B-P_A}{P_B(1-P_A-P_B)} + N_{AO} \cdot \frac{1}{1-P_A-P_B}$$

$$\Rightarrow (2N_{AA} + N_{AB} + N_{AO})(1-P_A-P_B) = (2N_O + N_{BO} + N_{AO}) P_A$$

$$(2N_{BB} + N_{AB} + N_{AO})(1-P_A-P_B) = (2N_O + N_{AO} + N_{BO}) P_B$$

$$\Rightarrow a(1-P_A-P_B) = b P_A$$

$$c(1-P_A-P_B) = b P_B$$

$$\Rightarrow \hat{P}_A = \frac{a}{a+b+c} \qquad \hat{P}_B = \frac{b}{a+b+c}$$

since $a+b+c = 2N_{AA} + 2N_{BB} + 2N_{AB} + 2N_{AO} + 2N_O = 2n$

we have

$$\hat{P}_A = \frac{2N_{AA} + N_{AB} + N_{AO}}{2n}$$

On the other hand, since this is multinomial distribution, the parameter estimats (MLE) for each of the six possibilities are just their counts devided by n, for example $\hat{P}_A^2 = \frac{N_{AA}}{n}$

So, since we have the constrain, and the hint, that leads to the solution

5. Another example of the EM algorithm. In genetics, we often assume a Hardy-Weinberg equilibrium. Consider the example of ABO blood type and read the following table. This is essentially saying that among human beings, the overall probability of having a Gene "A" is p_A, the probability of having Gene "B" is P_B, and the probability of having Gene "O" is p_O. Since each person will have two copies of a gene, one from each patient, it is possible that you have six different possible "Genotypes". Of course, the probability of having "AA" would be p_A^2, meaning that both your patients contribute a type "A" gene randomly to you. And the rest of the table should be self-explanatory. On the other hand, we know that your blood type (phenotype) will only be "A", "B", "O" and "AB", because, for example, if you have one "A" gene and one "O" gene, your blood type will be "A". And the rest of the possibilities are also demonstrated in the table.

Genotype	Probability	Phenotype
AA	P_A^2	A
AO	$2P_AP_O$	A
BB	P_B^2	B
BO	$2P_BP_O$	B
OO	P_O^2	O
AB	$2P_AP_B$	AB

Here comes the problem: we observe the phenotype (your blood type), but not the genotype (the types of your two genes) if without a gene sequencing machine. Hence, from the observed phenotype, it is difficult to estimate the probabilities P_A, P_B and P_O $(= 1 - P_A - P_B)$. An EM algorithm can be used. Suppose we observe n subjects and their phenotypes, we want to estimate these parameters. Then by the likelihood of a multinomial distribution, we should have the likelihood as (after removing some constants that involves combinatorial numbers)

$$L(P_A, P_B) = (P_A^2 + 2P_AP_O)^{n_A} \times (P_B^2 + 2P_BP_O)^{n_B} \times (P_O^2)^{(n_O)} \times (2P_AP_B)^{(n_{AB})}.$$

Here, n_A, n_B, n_AB, and n_O represent the number of subjects with the respective phenotype (blood type), and they should, of course, sum up to n. It is difficult to directly optimize this likelihood and solve for the parameters. However, we could introduce hidden variables n_{AO}, n_{AA}, n_{BO}, n_{BB}, which indicate the count of unobserved genotypes, respectively. And they should satisfy $n_{AO} + n_{AA} = n_A$ and $n_{BO} + n_{BB} = n_B$, hence we only need n_{AA} and n_{BB} as the hidden variables. The rest of the job is to perform the EM algorithm.

(a) (5 points) Write down the complete data log-likelihood (you can ignore unnecessary constants if you want) based on both the observed (n_A, n_B, n_{AB}, n_O) and unobserved (n_{AA}, n_{BB}) data. In other words, the analog of $\ell(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ (if you use our notation) of this problem.

(b) (5 points) For the E Step, calculate the conditional expectation of n_{AA} and n_{BB} given the observed data and the current parameter values. This is the analog of $E[Z|\mathbf{x}, \boldsymbol{\theta}]$. Hint: what is the conditional distribution of n_{AA} given n_A with all the other parameters known? This could be just one line of equation, but you should provide an explanation.

(c) (5 points) For the M Step, provide the formula to update the parameter estimate of P_A, given that all the counts are observed. Hint: from the relationship, we can realize that $2P_A^2 + 2P_AP_O + 2P_AP_B = 2P_A$. On the left-hand side, each quantity has its respective observed counts, on the right-hand side, it is our parameter of interest. This could be just one line of equation, but you should provide an explanation.

6. Variance of two-dimensional kernel density estimator. In HW4, we derived the bias part of a two-dimensional kernel density estimator. Let's finish the consistency proof by deriving the variance part (at a single target point). All assumptions of this question are exactly the same as HW4.

   (a) (8 points) Derive the variance of the estimator

   $$\mathrm{Var}[\widehat{f}(x)] =$$

   (b) (2 points) What is the rate of variance?

7. In the support vector machine problem, in some cases, we do not want to treating each observation equally. This means that we will assign a weight w_i to each observation. If an observation has a nonzero slack variable, we will penalize it proportional to the weights. This results in the following primal form of SVM in the non-separable case:

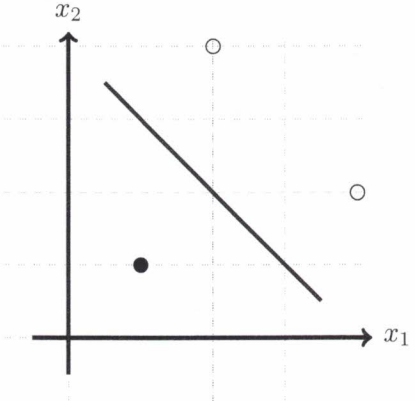
$$\begin{aligned} \min \quad & \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + C\sum_{i=1}^{n} w_i\xi_i \\ \text{subject to} \quad & y_i(x_i^T\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ & \xi_i \geq 0 \quad i = 1, \ldots, n \end{aligned}$$

   (a) (3 points) Write down the Lagrangian $\mathcal{L}$ of this constrained optimization problem by introducing two sets of Lagrange multiplier. It is recommended that you use α_i's and γ_i's.

   (b) (5 points) Optimize the Lagrangian by minimizing over $\boldsymbol{\beta}$ and β_0 and derive their solutions.

   (c) (5 points) Plug the solution back to the Lagrangian and find the dual form of this SVM problem.

   (d) (2 points) Suppose there are three observations (their labels are represented by dot and circle). Based on their current weights, the SVM decision line is plotted below. If the weights of all circles increase slightly, where do you expect the new decision line lies? Draw the new decision line approximately on the plot.



$$\mathrm{Var}[\hat{f}(x)] = \frac{1}{n}\underbrace{E\left[\frac{1}{\lambda^2}K^2\left(\frac{z-z_i}{\lambda}\right)\frac{1}{\lambda^2}K^2\left(\frac{w-w_i}{\lambda}\right)\right]}_{①} - \frac{1}{n}\underbrace{E\left[\frac{1}{\lambda}K\left(\frac{z-z_i}{\lambda}\right)\right]^2 \cdot E\left[\frac{1}{\lambda}K\left(\frac{w-w_i}{\lambda}\right)\right]}_{②}$$

$$① = \iint \frac{1}{\lambda^4}K^2\left(\frac{z-z_i}{\lambda}\right)K^2\left(\frac{w-w_i}{\lambda}\right)\cdot f(z_i, w_i)\,dz_i\,dw_i$$

$$= \iint \frac{1}{\lambda^2}K^2(t_1)K^2(t_2)\cdot f(z-t_1\lambda, w-t_2\lambda)\,dt_1 dt_2$$

$$= \iint \frac{1}{\lambda^2}K^2(t_1)K^2(t_1)\left[f(z,w) + \lambda\begin{pmatrix}t_1\\t_2\end{pmatrix}\nabla f(z,w) + \cdots\right]dt_1 dt_2$$

$$= \frac{1}{\lambda^2}\cdot C + o\left(\frac{1}{\lambda^2}\right)$$

$$② = O(1)$$

$$\frac{1}{n}① + \frac{1}{n}② \approx \frac{C}{n\lambda^2}$$

Hence the rate is $\frac{1}{n\lambda^2}$

Lagrangian:

(a) $L = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n} w_i\xi_i - \sum_{i=1}^{n}\alpha_i(y_i(x_i^T\beta+\beta_0) - (1-\xi_i)) - \sum_{i=1}^{n}\gamma_i\xi_i$, $\alpha_i, \gamma_i \geq 0$

(b)
$$\begin{cases} \frac{\partial L}{\partial \beta} = \beta - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \\ \frac{\partial L}{\partial \beta_0} = \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \frac{\partial L}{\partial \xi_i} = w_i C - \alpha_i - \gamma_i = 0. \end{cases} \Rightarrow \begin{cases} \beta = \sum_{i=1}^{n}\alpha_i y_i x_i \\ \sum_{i=1}^{n}\alpha_i y_i = 0 \\ w_i C - \alpha_i - \gamma_i = 0 \end{cases}$$

(c) plug back to L.

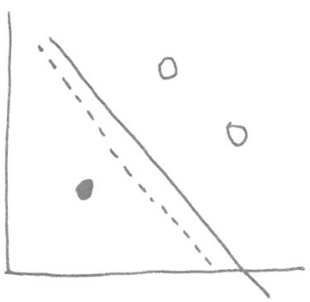
$$\max_{\alpha_i, \gamma_i} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0.$$

since $\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow w_i C - \alpha_i - \gamma_i = 0$ and $C > 0$, $\gamma_i \geq 0$, $\alpha_i > 0$

$$0 \leq \alpha_i \leq C \cdot w_i \quad , i = 1, \sim, n.$$

(d) when C is extremely large like ∞, this will not change the solution because any $\xi_i > 0$ is not beneficial, so ξ_i has to be exactly 0 making w_i useless. $C=0$ doesn't help either.



when C is small, moving the line toward dots receives less penalty than moving towards circle so the new line will move down slightly.