FINAL EXAMINATION–2021: SOLUTIONS

ORFE/FIN 525: Statistical Foundations of Data Science

May 15, 2021

1. cf codes.

2. (a) "Curse of dimensionality" is a common name for a deteriorating behavior of some statistical procedures once the dimension becomes large. It usually refers to nonparametric models whose effective number of parameter growths exponentially with dimensionality. The examples of models that AVOID the curse of dimensionality are
   - Additive model
   - Bivariate interaction models
   - Naive-Bayes type of models: using product of marginal densities as proxy of joint density
   - Neural network models

   (b) Yes, the k-means algorithm is the limit of EM-algorithm for Gaussian mixture model when $\mathbf{\Sigma}_k = \sigma^2\mathbf{I}$ for all clusters and $\sigma^2 \to 0$. The hidden variables in this case describe the assignment of points to clusters. The update step becomes the M-step and the assignment step becomes E-step. (See p. 13 of the presentation 6.)

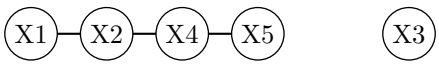
   (c) The answers are:
   - $(500+1)\times 200 + (200+1)\times 100 + (100+1)\times 200 + (200+1)\times 6 = 141706$ parameters (note that dropout does not add parameters).
   - 3 hidden layers (we do not count input and output),
   - 6 nodes in the output layer,
   - for multiclass classification (softmax activation of the last layer).

   (d) Preconditioning is a way to accelerate SGD by changing local geometry of the problem, making it well-conditioned. RMSprop uses the preconditioner that accumulates the information about the past gradients by exponential averaging, and this allows to facilitate fast convergence compared to SGD. (See p. 682 of the book.)

   (e) The factors are separated from the idiosyncratic part by taking first K principal components of the covariance input computed from the data. The crucial assumption is pervasiveness: due to this condition the top K eigenvalues are of order p and the rest are $O(1)$; hence the information contained in the principal components of covariance is almost not corrupted by the idiosyncratic part. (See p. 482 of the book.)

   (f) The factor adjustment is important when the features are highly correlated. In this case, standard model selection methods fail to control the number of selected features. In contrast, by including extracted factors we decorrelate and make model selection methods like SCAD work better. (See p. 5-6 of the presentation 10.)

   (g) It describes the conditional independence: $\omega_{ij} = 0$ if and only if X_i and X_j are conditionally independent given the rest variables. (See p. 452 of the book.) The corresponding graph:



3. (a) Since $\widehat{f}_n \in \mathcal{F}$, we have

$$R(\widehat{f}_n) - R_n(\widehat{f}_n) \leq \sup_{f \in \mathcal{F}} |R_n(f) - R(f)| = \Delta_n,$$

which implies the desired.

(b) From (a), $R(\widehat{f}_n) \leq R_n(\widehat{f}_n) + \Delta_n$. Similarly to (a), since $f^* \in \mathcal{F}$, we get $R_n(f^*) \leq R(f^*) + \Delta_n$. By definition of $\widehat{f}_n$, we have $R_n(\widehat{f}_n) \leq R_n(f^*)$. Combining these three inequalities, we get the desired:

$$R(\widehat{f}_n) \leq R_n(\widehat{f}_n) + \Delta_n \leq R_n(f^*) + \Delta_n \leq R(f^*) + 2\Delta_n.$$

(c) In this case, $R(f) = EI(f(\mathbf{X}) \neq Y) = P(f(\mathbf{X}) \neq Y)$, i.e. the probability of error of f, and $R_n(f) = n^{-1} \sum_{i=1}^n I(f(\mathbf{X}_i) \neq Y_i)$, i.e. the proportion of errors of f on the dataset.

(d) Note that

$$(Y - \mathbf{X}^\top \boldsymbol{\beta})^2 = \left[\begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}^\top \begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \right]^2 = \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}^\top \begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix}^\top \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}.$$

Thus, for $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$

$$R(f) = E(Y - \mathbf{X}^\top \boldsymbol{\beta})^2 = \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}^\top E \left[\begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix}^\top \right] \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}^\top \boldsymbol{\Sigma} \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix},$$

and

$$R_n(f) = n^{-1} \sum_{i=1}^n (Y_i - \mathbf{X}_i^\top \boldsymbol{\beta})^2 = \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}^\top n^{-1} \sum_{i=1}^n \left[\begin{pmatrix} \mathbf{X}_i \\ Y_i \end{pmatrix} \begin{pmatrix} \mathbf{X}_i \\ Y_i \end{pmatrix}^\top \right] \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}^\top \mathbf{S}_n \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}.$$

Hence,

$$|R_n(f) - R(f)| = \left| \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix}^\top (\mathbf{S}_n - \boldsymbol{\Sigma}) \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix} \right| \leq \|\mathbf{S}_n - \boldsymbol{\Sigma}\|_{\max} \left\| \begin{pmatrix} -\boldsymbol{\beta} \\ 1 \end{pmatrix} \right\|_1^2 = \|\mathbf{S}_n - \boldsymbol{\Sigma}\|_{\max} (1 + \|\boldsymbol{\beta}\|_1)^2.$$

Taking supremum over $f \in \mathcal{F}$ and taking into account that for such f holds $\|\boldsymbol{\beta}\|_1 \leq c$, we get the desired.

4. (a) Use the data $\{(\widehat{\delta}_1(\mathbf{x}_i), \ldots, \widehat{\delta}_m(\mathbf{x}_i))\}$ as covariates and Y_i as response ($i = 1, \ldots, n$) and fit the logistic regression model to obtain the coefficients $\{\widehat{\beta}_j\}_{j=0}^m$. The final ensemble classifier is simply

$$\widehat{\delta}_{ensemble}(\mathbf{x}) := I\left(\widehat{\beta}_0 + \sum_{j=1}^m \widehat{\beta}_j \widehat{\delta}_j(\mathbf{x}) > 0 \right).$$

(b) Using independence of $\widehat{\boldsymbol{\mu}}_d$ and $\widehat{\boldsymbol{\mu}}_a$,

$$E\widehat{\boldsymbol{\mu}}_d^\top (\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1) = (E\widehat{\boldsymbol{\mu}}_d)^\top (E\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \left(\frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} - \boldsymbol{\mu}_1 \right) = -\frac{\|\boldsymbol{\mu}_d\|^2}{2}.$$

Next,

$$E\|\widehat{\boldsymbol{\mu}}_d\|^2 = \|\boldsymbol{\mu}_d\|^2 + E\|\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|^2,$$

where the cross term disappears since $E\widehat{\boldsymbol{\mu}}_d = \boldsymbol{\mu}_d$. Further, denoting $\widehat{\boldsymbol{\mu}}_0$ and $\widehat{\boldsymbol{\mu}}_1$ the sample means of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, respectively, we get

$$E\|\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|^2 = E\|\widehat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0\|^2 + E\|\widehat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\|^2,$$

where the cross term again disappears due to independence of two samples. Finally,

$$\begin{aligned} E\|\widehat{\boldsymbol{\mu}}_0 - \boldsymbol{\mu}_0\|^2 &= E\left\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_0)\right\|^2 = \frac{1}{n}E\left\|\mathbf{X}_1 - \boldsymbol{\mu}_0\right\|^2 \\ &= \frac{1}{n}E\,\mathrm{tr}\left[(\mathbf{X}_1 - \boldsymbol{\mu}_0)(\mathbf{X}_1 - \boldsymbol{\mu}_0)^\top\right] = \frac{1}{n}\,\mathrm{tr}\,\mathbf{I}_p = \frac{p}{n}. \end{aligned}$$

Similarly, $E\|\widehat{\boldsymbol{\mu}}_1 - \boldsymbol{\mu}_1\|^2 = \frac{p}{n}$. Putting this all together,

$$E\|\widehat{\boldsymbol{\mu}}_d\|^2 = \|\boldsymbol{\mu}_d\|^2 + 2p/n.$$

This implies the desired:

$$\frac{E\widehat{\boldsymbol{\mu}}_d^\top(\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1)}{[E\|\widehat{\boldsymbol{\mu}}_d\|^2]^{1/2}} = -\frac{|\boldsymbol{\mu}_d\|^2/2}{[\|\boldsymbol{\mu}_d\|^2 + 2p/n]^{1/2}}.$$

Bonus part. Note that

$$\widehat{\boldsymbol{\mu}}_d^\top(\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1) = (\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d)^\top(\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_d^\top(\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_a) - \|\boldsymbol{\mu}_d\|^2/2.$$

By the Cauchy-Schwartz inequality, the expected value of first term is bounded by

$$(E\|\widehat{\boldsymbol{\mu}}_d - \boldsymbol{\mu}_d\|^2)^{1/2}(E\|\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1\|^2)^{1/2} = O(p/n)$$

by using what have already computed and the expected value of the second term is

$$\|\boldsymbol{\mu}_d\|(E\|\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1\|^2)^{1/2} = \|\boldsymbol{\mu}_d\|O(\sqrt{p/n})$$

Therefore, we have

$$\frac{E\widehat{\boldsymbol{\mu}}_d^\top(\widehat{\boldsymbol{\mu}}_a - \boldsymbol{\mu}_1)}{\|\boldsymbol{\mu}_d\|^2} = O_p\left(\frac{p}{n\|\boldsymbol{\mu}_d\|^2} + \frac{\sqrt{p}}{\sqrt{n}\|\boldsymbol{\mu}_d\|}\right) - \frac{1}{2}.$$

Therefore, the required condition is $\frac{p}{n\|\boldsymbol{\mu}_d\|^2} \to 0$.

(c) Note that the event $f_1(\mathbf{X}) > f_0(\mathbf{X})$ is identical to the event $p^{-1}\log(f_1(\mathbf{X})/f_0(\mathbf{X})) > 0$. Let us elaborate on $p^{-1}\log(f_1(\mathbf{X})/f_0(\mathbf{X}))$:

$$\frac{1}{p}\log\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} = \frac{1}{p}\log\frac{\prod_{k=1}^p g_1(X_k)}{\prod_{k=1}^p g_0(X_k)} = \frac{1}{p}\sum_{k=1}^{p}\log\frac{g_1(X_k)}{g_0(X_k)}.$$

Introducing for convenience the standardized version

$$z_k := \sigma_p^{-1}\left(\log\left(\frac{g_1(X_k)}{g_0(X_k)}\right) - \mu_p\right),$$

satisfying $E(z_k) = 0$, $\mathrm{var}(z_k) = 1$, we rewrite

$$\frac{1}{p}\log\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} = \mu_p + \frac{\sigma_p}{p}\sum_{k=1}^{p} z_k.$$

Therefore,

$$P\left(f_1(\mathbf{X}) > f_0(\mathbf{X})\right) = P\left(\frac{1}{p}\log\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > 0\right) = P\left(\mu_p + \frac{\sigma_p}{p}\sum_{k=1}^{p} z_k > 0\right) = P\left(\frac{1}{\sqrt{p}}\sum_{k=1}^{p} z_k > -\frac{\sqrt{p}\mu_p}{\sigma_p}\right).$$

If z_k satisfies regularity conditions required for CLT, we have

$$P\left(f_1(\mathbf{X}) > f_0(\mathbf{X})\right) \approx 1 - \Phi\left(-\frac{\sqrt{p}\mu_p}{\sigma_p}\right) = \Phi\left(\frac{\sqrt{p}\mu_p}{\sigma_p}\right)$$

as $p \to \infty$.

(d) **First solution**: Denote for shortness $a(x) = f_0(\mathbf{x})/(f_0(\mathbf{x})+g(\mathbf{x}))$. Then we rewrite our functional as

$$\begin{aligned} Q(g) &:= \int \left[f_0(\mathbf{x}) \log \frac{f_0(\mathbf{x})}{f_0(\mathbf{x}) + g(\mathbf{x})} + g(\mathbf{x}) \log \frac{g(\mathbf{x})}{f_0(\mathbf{x}) + g(\mathbf{x}))} \right] d\mathbf{x} \\ &= \int \left[(f_0(\mathbf{x}) + g(\mathbf{x}))(a(\mathbf{x}) \log(a(\mathbf{x})) + (1 - a(\mathbf{x})) \log(1 - a(\mathbf{x})) \right] d\mathbf{x}. \end{aligned}$$

By Jensen's inequality for convex function $x \log(x)$ we have

$$a(\mathbf{x}) \log(a(\mathbf{x})) + (1 - a(\mathbf{x})) \log(1 - a(\mathbf{x})) \geq \log(1/2),$$

implying $Q(g) \geq 2\log(1/2)$ for any g, since f_0 and g integrate to 1. It is straightforward to verify that this lower bound is attained exactly for $g(\mathbf{x}) = f_0(\mathbf{x})$.

Second solution: By the Lagrange method, the objective function becomes

$$\int \left[f_0(\mathbf{x}) \log \frac{f_0(\mathbf{x})}{f_0(\mathbf{x}) + g(\mathbf{x})} + g(\mathbf{x}) \log \frac{g(\mathbf{x})}{f_0(\mathbf{x}) + g(\mathbf{x}))} \right] d\mathbf{x} + \lambda \int g(\mathbf{x}) d\mathbf{x}.$$

Writing the summation of two integrals as one integral, the optimization problem is the same as minimizing the integrand for each given $\mathbf{x}$. Taking derivative with respect to $g(\mathbf{x})$ and set it to zero, we get

$$\frac{g(\mathbf{x})}{f_0(\mathbf{x}) + g(\mathbf{x})} + \lambda = 0,$$

which implies $g(\mathbf{x}) = c f_0(\mathbf{x})$ for a constant c. This constant must be 1, since $g(\mathbf{x})$ is a density.

5. (a) We assume $E\mathbf{f} = 0$ and hence $E\mathbf{X} = 0$. Then, using independence of $\mathbf{X}_1$ and $\mathbf{X}_2$

$$0.5E(\mathbf{X}_1 - \mathbf{X}_2)(\mathbf{X}_1 - \mathbf{X}_2)^\top = 0.5E\mathbf{X}_1\mathbf{X}_1^\top - 0.5E\mathbf{X}_1\mathbf{X}_2^\top - 0.5E\mathbf{X}_2\mathbf{X}_1^\top + 0.5E\mathbf{X}_2\mathbf{X}_2^\top = \mathbf{\Sigma}.$$

The (1,2)-th entry of robust U-statistic is

$$[\widehat{\mathbf{\Sigma}}_U]_{1,2} = \frac{1}{2\binom{n}{2}} \sum_{i \neq j} \min\left(1, \frac{\tau}{\|\mathbf{X}_i - \mathbf{X}_j\|^2}\right) (X_i^{(1)} - X_j^{(1)})(X_i^{(2)} - X_j^{(2)}),$$

where e.g. $X_i^{(1)}$ denotes the first entry of $\mathbf{X}_i$.

(b) Denoting $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_K$ and $\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_K$ to be the top K eigenvalues and eigenvectors of $\widehat{\mathbf{\Sigma}}_U$, respectively, we estimate $\mathbf{B}$ as

$$\widehat{\mathbf{B}} = [\widehat{\lambda}_1^{1/2}\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\lambda}_K^{1/2}\widehat{\boldsymbol{\xi}}_K].$$

(c) Note that

$$\begin{aligned} \|\widehat{\mathbf{B}}\mathbf{O} - \mathbf{B}\|_F^2 &= \|\widehat{\mathbf{B}}\|_F^2 + \|\mathbf{B}\|_F^2 - 2\,\mathrm{tr}(\mathbf{O}^\top \widehat{\mathbf{B}}^\top \mathbf{B}) \\ &= C - 2\,\mathrm{tr}(\mathbf{O}^\top \mathbf{L}\mathbf{\Lambda}\mathbf{R}) \end{aligned}$$

where C is a constant that does not depend on $\mathbf{O}$. Essentially, we need to maximize $\mathrm{tr}(\mathbf{O}^\top \mathbf{L}\mathbf{\Lambda}\mathbf{R})$ subject to $\mathbf{O}^\top\mathbf{O} = \mathbf{I}_K$.

First proof: Let $\widetilde{\mathbf{O}} = \mathbf{L}^\top\mathbf{O}\mathbf{R}^\top$. Then, it is also rotation matrix and the object becomes maximizing $\mathrm{tr}(\widetilde{\mathbf{O}}^\top\mathbf{\Lambda})$. By orthonormality, it is easy to see that $\mathrm{tr}(\widetilde{\mathbf{O}}^T\mathbf{\Lambda}) \leq \lambda_1 + \cdots + \lambda_K$ and the maximum is obtained when $\widetilde{\mathbf{O}} = \mathbf{I}_K$ and $\widehat{\mathbf{O}} = \mathbf{L}\mathbf{R}$.

Second proof: Using Lagrange multipliers $\mathbf{D}$ (w.l.o.g. we can assume it is symmetric), we form the Lagrangian

$$\ell(\mathbf{O}; \mathbf{D}) = \mathrm{tr}(\mathbf{O}^\top \mathbf{L}\mathbf{\Lambda}\mathbf{R}) + \mathrm{tr}(\mathbf{D}(\mathbf{O}^\top\mathbf{O} - \mathbf{I}_K))$$

and derive the first order condition

$$\frac{\partial \ell}{\partial \mathbf{O}} = \mathbf{L}\mathbf{\Lambda}\mathbf{R} + 2\mathbf{O}\mathbf{D} = 0.$$

Next, we express $\mathbf{D} = -\mathbf{O}^\top \mathbf{L}\mathbf{\Lambda}\mathbf{R}/2$ and compute

$$\mathbf{D}^2 = \mathbf{D}^\top \mathbf{D} = \frac{\mathbf{R}^\top \mathbf{\Lambda} \mathbf{L}^\top \mathbf{L} \mathbf{\Lambda} \mathbf{R}}{4} = \frac{\mathbf{R}^\top \mathbf{\Lambda}^2 \mathbf{R}}{4}.$$

Hence $\mathbf{D} = \pm \mathbf{R}^\top \mathbf{\Lambda} \mathbf{R}/2$, its inverse $\mathbf{D}^{-1} = \pm 2 \mathbf{R}^\top \mathbf{\Lambda}^{-1} \mathbf{R}$, which leads to

$$\widehat{\mathbf{O}} = -\frac{1}{2} \mathbf{L}\mathbf{\Lambda}\mathbf{R}\mathbf{D}^{-1} = \mp \mathbf{L}\mathbf{R},$$

where the lower sign corresponds to maximization and the upper sign to minimization. Since we are maximizing trace (to minimize Frobenius distance), we pick $\widehat{\mathbf{O}} = \mathbf{L}\mathbf{R}$.

(d) Denote $q_n = P(G(n,p) \text{ is not connected})$. Note that a graph is not connected if and only if there exist a partition of the set of nodes into two sets of sizes k and $(n-k)$ for some k (with $1 \le k \le n/2$) such that there is no edges between these sets. Hence, by union bound,

$$q_n \le \sum_{k=1}^{[n/2]} \binom{n}{k} (1-p)^{k(n-k)},$$

where the summation takes into account all possible k, factor $\binom{n}{k}$ stands for all possible options to split nodes into two sets of size k and $(n-k)$, and $(1-p)^{k(n-k)}$ is the probability that there in no edges between these sets.

Using standard inequalities $\binom{n}{k} \le n^k$ and and $\log(1-p) \le -p$, and plugging in $p = a\log(n)/n$ we get

$$q_n \le \sum_{k=1}^{[n/2]} \exp\left(\left(1 - \frac{a(n-k)}{n}\right) k \log(n) \right) \le \sum_{k=1}^{[n/2]} \exp\left((1 - a/2) k \log(n) \right).$$

If $a > 2$, we have $c := 1 - a/2 < 0$. By using the geometric series formula, we have

$$q_n \le \sum_{k=1}^{[n/2]} e^{ck\log(n)} = \sum_{k=1}^{[n/2]} z_n^k = \frac{1 - z_n}{1 - z_n^{[n/2]}} z_n,$$

where $z_n = e^{c\log(n)} = n^c$. When $n \to \infty$, we get $z_n \to 0$ and $q_n \to 0$.

If $2 \ge a > 1$, then we define $k^* := [n - 2n/(a+1)] \le n/2$ and split our sum into two parts

$$q_n \le \sum_{k=1}^{k^*} \binom{n}{k} (1-p)^{k(n-k)} + \sum_{k=k^*+1}^{[n/2]} \binom{n}{k} (1-p)^{k(n-k)}.$$

The first term is bounded similarly to the case $a > 2$ (but this time we need to take $c := (1-a)/2 < 0$), while for the second term we use

$$\sum_{k=k^*+1}^{[n/2]} \binom{n}{k} (1-p)^{k(n-k)} \le n \cdot 2^n e^{-pk^*(n-k^*)} = \exp\left(\log(n) + n\log(2) - \frac{a(a-1)}{a+1} n \log(n) \right) \to 0$$

as $n \to \infty$.

Expected number of connections for a given node is simply $(n-1)p$.