Final Exam

Quesion 1 (warming-up) (25 points)
Provide *short* (but *comprehensive*) answers to a series of brief questions:

1. Describe the problems of choosing "too small" and "too large" bandwidth in kernel estimation and local regression in terms of bias and variance of the resulting estimator. Is the residual sum of squares a good criterion for choosing a bandwidth? What will be the results of such a choice?

2. Which of the following (univariate) nonparametric regression estimators are linear and which are not: kernel estimator, local regression, smoothing spline, truncated generalized Fourier series, thresholded generalized Fourier series?

3. Which of the classification methods (LDA, QDA, logistic regression, k-NN, classification trees, neural networks, SVM) face conceptual problems whith categorical explanatory variables? Explain (briefly) why.

4. If the true boundaries between classes are linear, do you expect LDA or QDA to perform better on the training set? What about the test set? What happens if the true boundaries are nonlinear?

5. In using CART can one *always* grow a sufficiently large tree to achieve zero impurity measure at each terminal node?

6. Consider a binary classification with a *single* explanatory continuous variable x. What will be the form of the decision boundary for the LDA and logistic regression in this case? Will these conclusions necessarily remain true for CART and for the k-NN (say, 1-NN for simplicity)? (if *yes* - explain, if *no* - give (draw?) a counterexample)

7. Find the architecture of the neural network that mimics the multinomial logistic regression.

8. Find the VC-dimension of the family of spherical classifiers in $\mathbb{R}^2$, that is $\{\eta(\mathbf{x}) : \eta(\mathbf{x}) = I\{\mathbf{x} \in S_{\mathbf{a},r}\}, \mathbf{a} \in \mathbb{R}^2, r > 0\}$, where $S_{\mathbf{a},r} = \{\mathbf{x} \in \mathbb{R}^2 : ||\mathbf{x} - \mathbf{a}||_2 \leq r\}$.

9. Show that the quadratic loss $\varphi(u) = (1-u)^2$ is calibrated.

Quesion 2 (15 points)
Consider a general nonparametric regression model:

$$y_i = g(x_i) + \epsilon_i, \quad i = 1, ..., n, \; E\epsilon = \mathbf{0}, Var(\epsilon) = \sigma^2 I_n$$

Consider a general linear estimator (smoother) $\hat{\mathbf{g}} = S\mathbf{y}$ with a matrix $S \in \mathbb{R}^{n \times n}$. If $y_i = c$ for all $i = 1, \ldots, n$, a reasonable linear estimator should evidently estimate g_i's by the same constant c, i.e. $\hat{\mathbf{g}} = c\mathbf{1}$ in this case.

1. Show that for an estimator with such a property the sum of elements for each row of S is one.

2. Show that $Cov(\mathbf{Y}, \hat{\mathbf{g}}) = E\left((\mathbf{Y} - E\mathbf{Y})^T(\hat{\mathbf{g}} - E\hat{\mathbf{g}})\right) = \sigma^2 tr(S)$.

3. For a general linear estimator $\hat{g}$ it is not always obvious what is meant by the corresponding "leave-one-out" estimator $\hat{g}_{(-i)}$. To keep the above property from 2.1 for $\hat{g}_{(-i)}$, define it as

$$\hat{g}_{(-i)}(x_i) = \frac{\sum_{j \neq i} S_{ij} y_j}{\sum_{j \neq i} S_{ij}}$$

Show that is this case the cross-validation

$$CV = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{g}_{(-i)}(x_i))^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{g}(x_i)}{1 - S_{ii}} \right)^2$$

Question 3 (35 points)
Consider the standard (univariate) nonparametric regression model

$$y_i = g(x_i) + \epsilon_i, \quad i = 1, ..., n,$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and σ^2 is assumed to be known.

Let's focus on the generalized Fourier series approach. Following this methodology, we expand g in a certain orthonormal basis $\{\psi_j(x)\}$ as $\sum_{j=1}^{\infty} \beta_j \psi_j(x)$, approximate g by truncated $g_n(x) = \sum_{j=1}^{n} \beta_j \psi_j(x)$ and estimate the resulting vector of generalized Fourier coefficients $\boldsymbol{\beta} \in \mathbb{R}^n$ from the empirical generalized Fourier coefficients $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^n$ obtained by discrete generalized Fourier transform of the data vector $\mathbf{Y}$. Suppose for simplicity that the corresponding discrete generalized Fourier transform is also orthonormal (like DFT or DWT for equidistant design).

1. What is the joint distribution of the vector $\tilde{\boldsymbol{\beta}}$?

2. Consider the following *testimation* procedure: first, test (simultaneously) all $\tilde{\beta}_j$, $j = 1, \ldots, n$ for significance by Bonferroni multiple testing at level α and then estimate β_j by $\tilde{\beta}_j$ if it was found significant and zero it otherwise. Show that it results in hard thresholding with a universal (not depending on the data) threshold. Show that for large n, it is similar to the universal thresholding of Donoho and Johnstone. *Hint*: you can use the approximation $z_{1-q} \sim \sqrt{2 \ln(1/q)}$ for small q.

3. As we have discussed in the class, to get consistent estimators of the unknown $\boldsymbol{\beta}$ one usually performs linear shrinkage or thresholding of the empirical coefficients $\tilde{\beta}_j$'s.

   (a) Consider first linear shrinkage estimators $\hat{\beta}_j = w_j \tilde{\beta}_j$ with given weights w_j.

      i. Find the average mean squared error $AMSE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = \frac{1}{n} E ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2$ of the above linear estimator.

      ii. Suppose that an oracle revealed you the real values of β_j's while you were dreaming at night but forbid you to tell anyone about it. What weights w_j will you choose in the morning to minimize AMSE? What will be the corresponding (ideal) AMSE?

   (b) Consider now thresholding.

      i. Still being excited by the information revealed by the oracle, you want to use it for choosing an optimal threshold λ: you threshold $\tilde{\beta}_j$ if the corresponding *true* $|\beta_j| < \lambda$ and keep $\tilde{\beta}_j$ as it is otherwise (hard thresholding) - recall that from the oracle you *do* know the true β_j's! What will be your choice for the optimal threshold λ and the resulting (ideal) AMSE?

ii. Unfortunately, at the end, it was just a dream, there was no oracle... So now you face the real thresholding problem when you do not know true β_j's and threshold the *observed* $\tilde{\beta}_j$'s by hard thresholding: $\hat{\beta}_j = \tilde{\beta}_j I\{|\tilde{\beta}_j| > \lambda\}$. For a given threshold λ derive the $MSE(\hat{\beta}_j, \beta_j) = E(\hat{\beta}_j - \beta_j)^2$ for a single coefficient. Try to get the expression as simple as possible but do not be desperate if it is still not so "nice" you are used to in the *linear* world and involves, for example, the cumulative distribution function $\Phi(\cdot)$ and the density function $\phi(\cdot)$ of the standard normal distribution (o-ps... I am afraid I have said too much so I would better stop here... (-:)).

Question 4 (25 points)
Suppose we have data $(\mathbf{X}_i, Y_i), \quad i = 1, \ldots, n$, where $\mathbf{X}_i$'s are d-dimensional vectors from two groups ($Y_i = 1$ or $Y_i = 2$). For the first group $E(\mathbf{X}|Y = 1) = \boldsymbol{\mu}_1$ and $Var(\mathbf{X}|Y = 1) = \Sigma$, while for the second one $E(\mathbf{X}|Y = 2) = \boldsymbol{\mu}_2$ and $Var(\mathbf{X}|Y = 2) = \Sigma$ (the covariance matrix is the same for both groups). We wish to find a linear combination $U = \mathbf{a}^T\mathbf{X}$ of the components of $\mathbf{X}$ (or, geometrically, the direction for the one-dimensional projection of the data) that maximizes the separation between the two groups in the sense that

$$\frac{(E(U|Y = 1) - E(U|Y = 2))^2}{Var(U|Y)} \to \max_{\mathbf{a}\in\mathbb{R}^d}$$

(or, equivalently, maximizes the ratio of the between-class variance to the within-class variance). In fact, this is the original criterion for discrimination between two groups proposed by Fisher.

1. What are $E(U|Y = 1)$, $E(U|Y = 2)$, $Var(U|Y = 1)$ and $Var(U|Y = 2)$?
2. Show that the optimal vector $\mathbf{a}_*$ is the eigenvector of the matrix $\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ and $\mathbf{a}_* = c\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ for any constant $c \neq 0$.
3. Define the midpoint $m_0 = \frac{1}{2}(\mathbf{a}_*^T\boldsymbol{\mu}_1 + \mathbf{a}_*^T\boldsymbol{\mu}_2)$ and naturally classify a new observation vector $\mathbf{x}_0$ to the first group iff $\mathbf{a}_*^T\mathbf{x}_0 > m_0$. What is the resulting classifying rule (in terms of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ)? Is it a linear classifier?
4. Is it related to the LDA classifier?

- The deadline is **Friday, 28 February, 12:00 noon**. Since it is Friday, send me your (*clearly written*) exam by email but if you can also put its (identical!) hard copy in my box on Sunday-Monday, I would be grateful.
- If something is not clear or you have questions, you can email $felix@tauex.tau.ac.il$.

Good Luck!