Projects for the course Advanced Statistical Methods

General project requirements

Each project aims to check the performance of a model selection procedure on a synthetic dataset. Data generation process and detailed project descriptions are given in the sections below. A project report must consist of a code and of a presentation containing all the necessary formulae and plots. You may use Jupyter Notebook to combine the presentation and the code in one file. You may use all implemented functions from the standard Python libraries (scikit-learn, numpy, scipy, etc.), so you do not have to implement them by yourself.

Data generation

You have an access to a sample $S_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$ where $X_i = i/n, i \in \{1, \dots, n\}$, are equidistant design points, and Y_i's are generated from the model

$$Y_i = f^*(X_i) + \varepsilon_i, \quad 1 \leq i \leq n, \tag{1}$$

where f^* is a univariate function on $[0,1]$, and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. standard Gaussian random variables $\mathcal{N}(0,1)$. In all the projects, the sample size n is set to 200 and the function f^* is generated artificially according to the following scheme. Consider the Fourier basis $\{\psi_j(x)\}_{j=1}^n$:

$$\psi_j(x) = \begin{cases} 1, \quad \text{if } j = 0, \\ \sin(\pi(j+1)x), \quad \text{if } j \text{ is odd}, \\ \cos(\pi j x), \quad \text{if } j \text{ is even}. \end{cases}$$

The true function f^* is then equal to

$$f(x) = c_1\psi_1(x) + \dots + c_n\psi_n(x),$$

where the coefficients $c_1, \dots, c_n$ are chosen randomly: with γ_j i.i.d. standard normal,

$$c_j = \begin{cases} \gamma_j, \quad 1 \leq j \leq 10, \\ \frac{\gamma_j}{(j-10)^2}, \quad 11 \leq j \leq n \end{cases}$$

The simulation must be reproducible, fix a seed when generate pseudo-random numbers. Provide the plot of the function f^*.

Remark 1. Though you know f^*, please, ensure that you do not use any knowledge of it when construct your estimates. You have an access to the sample $S_n = \{(X_i, Y_i) : 1 \le i \le n\}$ only.

Project 1: function estimation via SURE

The goal of the project is to estimate f^*. First of all, approximate f^* choosing a basis from the following options (a parameter to tune is given in brackets):

B1. piecewise constant (number of pieces);

B2. piecewise linear (number of pieces);

B3. natural cubic splines (number of knots);

B4. smoothing splines (penalty parameter).

After you chose a basis for approximation, fix a family $\mathscr{M}$ of parameters where you are going to find the best one. The set $\mathscr{M}$ should include at least 10 elements. For each $m \in \mathscr{M}$, compute the design matrix Ψ_m. Apply the unbiased risk estimation to adaptively choose a parameter $\widehat{m} \in \mathscr{M}$.

Evaluate the performance of the procedure. The performance of the final estimate is measured by

$$\mathcal{R}(\widehat{f}) = \mathbb{E}\|f^* - \widehat{f}\|^2 \equiv \sum_{i=1}^{n} \mathbb{E}(f^*(X_i) - \widehat{f}(X_i))^2.$$

For each model $m \in \mathscr{M}$, compute the true risk $\mathcal{R}_m$, corresponding to the maximum likelihood estimate for this model. Note that, for each $m \in \mathscr{M}$, the true risk is deterministic. Provide an explicit formula for computation of the true risk and a plot how it depends on the parameter. After you made the plot, print the oracle choice of the parameter $m^* \in \operatorname*{argmin}_{m \in \mathscr{M}} \mathcal{R}_m$ and the oracle risk $\mathcal{R}_{m^*}$. Compare the risk of your estimate $\mathcal{R}_{\widehat{m}}$ with $\mathcal{R}_{m^*}$. Compare the risk of your estimate $\mathcal{R}_{\widehat{m}}$ with $\mathcal{R}_{\widetilde{m}}$, where $\widetilde{m}$ is chosen by

CV1. 10-fold cross-validation;

CV2. leave-one-out cross-validation.

Remark 2. You can use f^* when compute the true risk.

Project 2: derivative estimation via SURE

The goal of the project is to estimate the derivative of f^*. First of all, approximate f^* choosing a basis from the following options (a parameter to tune is given in brackets):

B1. piecewise linear (number of pieces);

B2. natural cubic splines (number of knots);

B3. smoothing splines (penalty parameter).

After you chose a basis for approximation, fix a family $\mathscr{M}$ of parameters where you are going to find the best one. The set $\mathscr{M}$ should include at least 10 elements. For each $m \in \mathscr{M}$, compute the design matrix Ψ_m. For each model, find the corresponding MLE. Apply the unbiased risk estimation for linear functionals (Section 6 in the script) to adaptively choose a parameter $\widehat{m} \in \mathscr{M}$.

Evaluate the performance of the procedure. The performance of the final estimate is measured by

$$\mathcal{R}(\widehat{f}') = \mathbb{E}\|(f^*)' - \widehat{f}'\|^2 \equiv \sum_{i=1}^{n} \mathbb{E}((f^*)'(X_i) - \widehat{f}'(X_i))^2.$$

For each model $m \in \mathscr{M}$, compute the true risk $\mathcal{R}_m$, corresponding to the maximum likelihood estimate for this model. Note that, for each $m \in \mathscr{M}$, the true risk is deterministic. Provide an explicit formula for computation of the true risk and a plot how it depends on the parameter. After you made the plot, print the oracle choice of the parameter $m^* \in \operatorname*{argmin}_{m \in \mathscr{M}} \mathcal{R}_m$ and the oracle risk $\mathcal{R}_{m^*}$. Compare the risk of your estimate $\mathcal{R}_{\widehat{m}}$ with $\mathcal{R}_{m^*}$.

Remark 3. You can use f^* and $(f^*)'$

Project 3: pointwise function estimation via SURE

Draw a point x_0 from the uniform distribution on $[0, 1]$. The goal of the project is to estimate $f^*(x_0)$, applying kernel smoothing and unbiased risk estimation. Fix an decreasing sequence of positive numbers $\{h_m : 1 \leq m \leq M\}$, $M \geq 10$. For each bandwidth m, estimate $f^*(x_0)$ using one of the following estimates:

E1. locally constant estimate;

E2. locally linear estimate;

E3. locally quadratic estimate

and one of the following kernels:

K1. rectangular kernel $\mathcal{K}(x) = \mathbb{1}(|x| \leq 1)$;

K2. Epanechnikov kernel $\mathcal{K}(x) = (1 - x^2)_+$;

K3. Gaussian kernel $\mathcal{K}(x) = e^{-x^2/2}$.

Apply the unbiased risk estimation procedure for linear functionals (Section 6 in the script) to choose the best bandwidth amongst $\{h_m : 1 \le m \le M\}$. Note that the map $f \mapsto f(x_0)$ is a linear functional.

Evaluate the performance of the procedure. The performance of the final estimate is measured by

$$\mathcal{R}(\widehat{f}) = \mathbb{E}(f^*(x_0) - \widehat{f}(x_0))^2.$$

For each bandwidth $h_m, m \in \{1, \dots, M\}$, compute the true risk $\mathcal{R}_m$, corresponding to the local estimate for this model. Note that, for each $m \in \{1, \dots, M\}$, the true risk is deterministic. Provide an explicit formula for computation of the true risk and a plot how it depends on the parameter. After you made the plot, print the oracle choice of the parameter $m^* \in \operatorname*{argmin}_{1 \le m \le M} \mathcal{R}_m$ and the oracle risk $\mathcal{R}_{m^*}$. Compare the risk of your estimate $\mathcal{R}_{\widehat{m}}$ with $\mathcal{R}_{m^*}$.

Remark 4. You can use f^* and $(f^*)'$ when compute the true risk.

Project 4: pointwise derivative estimation via SURE

Draw a point x_0 from the uniform distribution on $[0, 1]$. The goal of the project is to estimate $(f^*)'(x_0)$, applying kernel smoothing and unbiased risk estimation. Fix a decreasing sequence of positive numbers $\{h_m : 1 \le m \le M\}$, $M \ge 10$. For each bandwidth m, estimate $f^*(x_0)$ using one of the following estimates:

E1. locally linear estimate;

E2. locally quadratic estimate;

E3. locally cubic estimate

and one of the following kernels:

K1. rectangular kernel $\mathcal{K}(x) = \mathbb{1}(|x| \le 1)$;

K2. Epanechnikov kernel $\mathcal{K}(x) = (1 - x^2)_+$;

K3. Gaussian kernel $\mathcal{K}(x) = e^{-x^2/2}$.

Apply the unbiased risk estimation procedure for linear functionals (Section 6 in the script) to choose the best bandwidth amongst $\{h_m : 1 \le m \le M\}$. Note that the map $f \mapsto f'(x_0)$ is a linear functional.

Evaluate the performance of the procedure. The performance of the final estimate is measured by

$$\mathcal{R}(\widehat{f}) = \mathbb{E}((f^*)'(x_0) - \widehat{f}'(x_0))^2.$$

For each bandwidth $h_m, m \in \{1, \dots, M\}$, compute the true risk $\mathcal{R}_m$, corresponding to the local estimate for this model. Note that, for each $m \in \{1, \dots, M\}$, the true risk is deterministic. Provide an explicit formula for computation of the true risk and a plot how it depends on the parameter. After you made the plot, print the oracle choice of the parameter $m^* \in \operatorname*{argmin}_{1 \le m \le M} \mathcal{R}_m$ and the oracle risk $\mathcal{R}_{m^*}$. Compare the risk of your estimate $\mathcal{R}_{\widehat{m}}$ with $\mathcal{R}_{m^*}$.

Remark 5. You can use f^* and $(f^*)'$ when compute the true risk.

Project 5: parameter estimation via full Bayes approach

Approximate f^* using one of the following bases:

B1. piecewise constant (number of pieces);

B2. piecewise linear (number of pieces);

B3. natural cubic splines (number of knots).

The number of pieces or the number of knots should be taken sufficiently large. Compute the best parametric fit:

$$\theta^* = \operatorname*{argmin}_{\theta \in \mathbb{R}^p} \mathbb{E} L(\theta) = (\Psi \Psi^\top)^{-1} \Psi Y.$$

The goal is to estimate θ^* using the Bayesian inference. Fix an increasing sequence of positive numbers $\{\tau_m : 1 \le m \le M\}$, $M \ge 10$, and consider a family of priors $\mathcal{N}(0, \tau_m^{-2} I_p)$ and a uniform hyper-prior on $\{\tau_1, \ldots, \tau_M\}$.

Perform the following tasks:

- Sample the model parameter τ_m, $1 \le m \le M$, from the hyper-prior and then the parameter θ_m, from the conditional prior $(\vartheta \,|\, \tau_m)$.
- Compute the corresponding hyper-posterior $\exp\{L(\theta, \tau, \,|\, Y)\}$.
- Compute and plot the marginal hyper-posterior $(\tau \,|\, Y)$ against the risk-based posterior.
- Compute and plot the marginal posterior for the target parameter $(\vartheta \,|\, Y)$.

Evaluate the performance of the estimate. Compute the frequentist risk $\mathcal{R} = \mathbb{E}\|\theta - \theta^*\|^2$, where $\theta \sim (\vartheta \,|\, Y)$. For each $m \in \{1, \ldots, M\}$, compute the frequentist risk $\mathcal{R}_m = \mathbb{E}\|\theta_m - \theta^*\|^2$, where $\theta_m \sim (\vartheta \,|\, \tau_m, Y)$. Plot the dependence $\mathcal{R}_m$ of m and find $m^* \in \operatorname*{argmin}_{1 \le m \le M} \mathcal{R}_m$. Compare $\mathcal{R}$ with the risk of the best model $\mathcal{R}_{m^*}$. Compare the risk $\mathcal{R}$ with $\mathcal{R}_{\widetilde{m}}$, where $\widetilde{m}$ is chosen by

CV1. 10-fold cross-validation;

CV2. leave-one-out cross-validation.