Problem sheet 2: Model assessment and selection

Analisi dei Dati – CdS Matematica – 2024/25

When suitable, please provide summmarising and explanatory pictures.

Exercise 2.1 (Local minima for the $0-1$ loss)**.** Construct a hypothesis class (parametrised by a finite number of parameters) and a n-uple S in $\mathbb{R}^2 \times \{-1, 1\}$ so that there are parameters θ_0, $\theta_\star$ and a value $\epsilon > 0$ such that

- for each θ such that $\|\theta - \theta_0\| \leqslant \epsilon$, $\mathcal{R}_S(\theta_0) \leqslant \mathcal{R}_S(\theta)$ (that is, the predictor associated to the parameter θ_0 is a local minimum),
- $\mathcal{R}_S(\theta_\star) < \mathcal{R}_S(\theta_0)$ (that is, θ_0 is not a global minimum).

Construct a (non-trivial) distribution $\mathcal{D}$ on $\mathbb{R}^2 \times \{-1, 1\}$ such that the same happens for samples S with distribution $\mathcal{D}$ with positive probability. Determine finally a distribution so that local minima exist with probability one.

Exercise 2.2. Consider the dataset `Ames`[1] from `R` package *modeldata*. Split the dataset in training and test subset, fit a linear regressive model with a arbitrarily chosen output (usually `Sale_Price`) and estimate the test error. Analyse the trend of test error when training and test set are split in different proportions. You may drop the categorical variables for simplicity, if you wish.

Exercise 2.3. Generate data, with polynomial dependence and perturbed by noise. Develop structured risk minimization to identify the optimal degree, in both cases of uniform and non-uniform weights.

Exercise 2.4 (Failure of CV)**.** Consider a binary classification problem with labels uniformly distributed on $\{0, 1\}$. Let h be the predictor that returns 1 if the sum of labels of the sample is odd, and 0 otherwise. Show that the difference, in absolute value, between the theoretical error and the LOO-CV error is $\frac{1}{2}$.

Exercise 2.5. Generate a dataset, apply a prediction technique (linear regression or k-nearest-neighbour, for instance) to evaluate statistically the test error.

- Evaluate the theoretical error (empirical mean and standard deviation).
- Evaluate the error through *hold out set* validation, LOO-CV and k-fold CV (empirical mean and standard deviation).
- Evaluate in particular the test error trend for different values of k.

Exercise 2.6. Consider again the dataset generated in Exercise 2.3. Identify the optimal degree through *hold out set* validation, LOO-CV, and k-fold CV. Compute training and validation error in dependence of the degree.

Exercise 2.7. Consider the dataset `Ames` from `R` package *modeldata* through linear regression.

- Compute training and test error of *hold out set* validation evaluated on random samples of the training/validation sets.
- Compute training and test error of LOO-CV.
- Compute training and test error of k-fold CV, for different values of k.

Exercise 2.8. Consider the dataset `Ames` from `R` package *modeldata* through linear regression (output of your choice). Find the *learning curve* with an estimate of the test error obtained through *hold out set* validation, LOO-CV, k-fold CV and bootstrap.

Exercise 2.9. Consider a sample of size n and a bootstrap sample B. Let D be the random variable that counts the number of elements of B when repetitions are neglected.

- Compute mean and variance of D.
- Find the distribution of D.

[1] Earlier versions of this set of exercises used the dataset `Boston` from package *MASS*. This turned out to be not a good idea from an ethical point of view, see for instance https://medium.com/@docintangible/racist-data-destruction-113e3eff54a8.