Machine Learning Part I (B. Wilbertz)
M2 Probabilités & Finance, UPMC-Ecole Polytechnique
M2 P.M.A.
12th January, 2024

1.5h, books and mobile phones not allowed

A Kaggle in-class competition

1. (*48 points*) State your team name for the competition (probably your student id or full name)
2. (*1 point*) Which algorithms did you use for the final submission?
3. (*1 point*) What was your biggest insight when participating in the competition?

B Evaluation of ML trainings

(*10 points*) In a binary classification problem (classes `A` and `B`) a machine learning classifier is producing the following results:

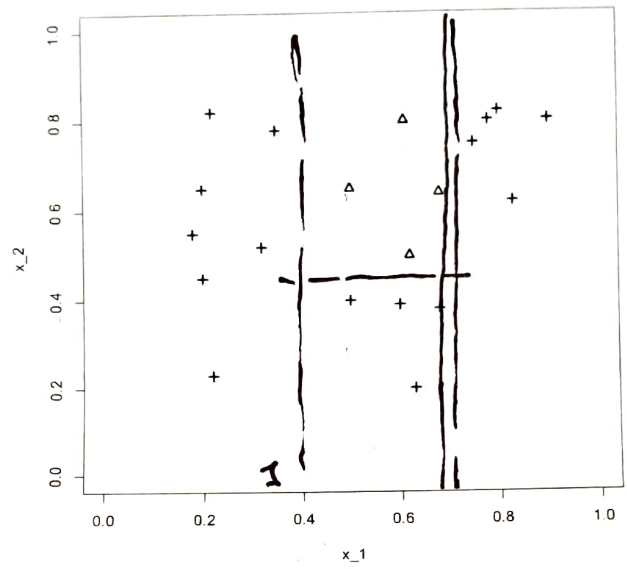
probs	ground truth
0.1	A
0.2	A
0.3	A
0.4	B
0.5	A
0.6	B
0.7	A
0.8	B
0.9	B
1.0	B

(probs are confidences for class `B`).

1. Draw the ROC curve for this model
2. Compute the AUC (area-under-the-curve) value for this model

C Decision Tree

(*10 points*) Consider the following binary classification problem with inputs (x_1, x_2) and outcome $y \in \{\triangle, +\}$.



1. *Compute the Gini index for this dataset*
2. Perform the splitting operation of the decision tree algorithm on this dataset using the Gini index as loss criterium (spliting values should be estimated from the plot). Report the Gini loss reduction for every split and continue until no further reduction of the loss is possible.
3. Draw the resulting classifier in tree form

D Causal Inference

(*15 points*) Consider the following scenario: Physical vulnerability (Z) has a direct effect on the likelihood of fatality (Y) and whether or not someone chooses to get vaccinated (X). Vaccination has a direct effect on the chances of fatality. In the entire population, 50% of the people are susceptible to a certain disease. For healthy and vaccinated people, the fatality rate is 1%. For healthy but unvaccinated people, the fatality rate is 4%. For vulnerable and vaccinated people, the fatality rate is 5.8%. For vulnerable and unvaccinated people, the fatality rate is 7%. Overall, the fatality rate for vaccinated people is 5%, while the fatality rate for unvaccinated people is 4.5%.

1. Draw the causal diagram for the effect of vaccination (X) on death (Y).
2. Compute the average causal effect of this vaccination on fatality i.e.

$$\mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$$

E Multiple choice questions

For each question only one possible answer is correct.

1. (*1 point*) What are typical numbers for K in K-fold cross-validation?
   - (a) $K = 3$ and $K = 9$
   - (b) $K = 4$ and $K = 8$
   - (c) $K = 4$ and $K = 10$
   - •(d) $K = 5$ and $K = 10$

2. (*1 point*) What is the usual relation between K in in K-fold cross validation and the number of bootstrap iterations n?
   - • (a) $n > K$
   - (b) $n \approx K$
   - (c) $n < K$
   - (d) n can be larger or smaller than K

3. (*1 point*) Which metric should be used for class imbalanced data sets?
   - (a) *Accuracy*
   - (b) *Specificity*
   - • (c) *AUC*
   - (d) *RMSE*

4. (*1 point*) Which of the following metrics does not make sense for regression problems?
   - (a) AIC
   - (b) R^2
   - (c) RMSE
   - • (d) ROC

5. (*2 points*) Which statement is correct?
   - • (a) The lasso method yields sparse models
   - (b) The ridge method yields sparse models
   - (c) The ridge method is designed to lower model bias
   - (d) The drawback of the lasso method is an increased model variance

6. (*2 points*) Which statement is correct?

   (a) SVM methods only works for linearly separable data

   (b) SVM method needs no cross-validation for hyperparameter tuning

   (c) The kernel trick for SVMs allows separation in a lower dimensional space

   • (d) The soft-margin criterion tolerates some violation of the data separation assumption

7. (*2 points*) Which of the following principles makes bagging work for random forests?

   • (a) Decorrelation of trees by choosing random predictor subsets

   (b) Pruning of the trees

   (c) Increasing the depth of the trees

   (d) Penalty term on the weights

8. (*1 point*) What is the typical number m of predictors to be taken at each split in the random forest algorithm for classification (total number of predictors p, size of training set n)?

   (a) $p/2$

   (b) $p/3$

   • (c) $\sqrt{p}$

   (d) $\log(p)$

9. (*2 points*) What can we say about the predictors when a gradient boosting algorithm yields only depth-1 trees?

   (a) We have to train with less data

   (b) We have to train with more data

   • (c) There is no interaction between the predictors

   (d) All predictors are linearly dependent

10. (*2 points*) Which statement is correct?

    • (a) Gradient boosting can overfit if the number of trees becomes too large

    (b) Random Forest can overfit if the number of trees becomes too large

    (c) A small learning rate η (also called shrinkage parameter) for gradient boosting reduces the number of trees needed

    (d) The boosting principle favors large trees to prevent overfitting

Examen Apprentissage Statistique
Partie 2 : Deep learning
Master Probabilités Finance – Sorbonne Université – Ecole Polytechnique
2024-01-12 - Durée 1h30 – Notes de cours papier autorisés

Questions GANs

Answer Y/N just before the question number please.

1. A GAN allows us to learn any distribution.
2. The main applications of GANs are for text generation.
3. Inference in GANs amounts at sampling from the latent space ($variable\ z$) and then computing the output of the generator in a deterministic way.
4. The training algorithm for GAN is a stochastic gradient descent algorithm used for optimizing an approximation of the data likelihood.
5. The GAN discriminator optimizes a cross-entropy criterion.
6. If for a given generator, the discriminator perfectly separates the ground truth data and the simulated data, the gradient w.r.t. the parameters of the discriminator is 0.
7. Let us suppose that the GAN has reached its equilibrium, then the optimal discriminator computes $D(x) = \frac{1}{2}, \forall x$.
8. The distribution of the latent variables (z) is usually chosen as a simple distribution (Gaussian, etc).
9. The objective of GAN is to generate a distribution that is indistinguishable from the distribution of the observed data..
10. For training the GANs one samples from the space of latent variables (z) and from the data space (x).

Exercise: the whole exercise is about ensemble methods.

We consider data from a source space $\mathcal{X}$ and label space $\mathcal{Y}$, $l: \mathcal{Y}^2 \to R^+$ a loss function, a sample is denoted $(x, y) \in \mathcal{X}x\mathcal{Y}$, and its distribution is denoted $p(x, y)$ or p for short. Let us consider a neural network $f(., \theta): \mathcal{X} \to \mathcal{Y}$, with parameters θ. The objective of training consists in minimizing the generalization error defined as $\mathcal{L}(f, \theta) = E_{(x,y)\sim p}[l(f(x, \theta), y)]$. For that the network will be trained on a dataset D of size N, sampled from p using a training configuration c that may include different sources of randomness such as hyperparameters. We denote $e = (D, c)$ the corresponding learning procedure, $\theta(e) = argmin_\theta \frac{1}{N} \sum_{(x,y)\in D} l(f(x, \theta), y)$ denotes the weights learned through e. The generalization error of functions learned under the learning procedure $e \sim p_e$, with p_e a distribution on e, is defined as $\mathcal{L}_e(f) = E_{e\sim p_e}[\mathcal{L}(f, \theta(e))]$. Being an expectation, $\mathcal{L}_e(f)$ does not depend on a particular training set D or configuration c. It characterizes the performance of the class of estimators learned through e. In the following we will consider real valued functions f and MSE loss i.e. $l(f(x, \theta), y) = (f(x, \theta) - y)^2$.

1. Show that $\mathcal{L}_e(f) = E_{e\sim p_e}[\mathcal{L}(f, \theta(e))] = \mathrm{E}_{(x,y)\sim p}[bias^2(f|(x, y)) + var(f|x)]$ (1)

With $bias(f|(x, y)) = y - \bar{f}(x), var(f|x) = E_e\left[\left(f(x, \theta(e)) - \bar{f}(x)\right)^2\right], \bar{f}(x) = E_e[f(x, \theta(e))]$

2. We now consider an ensemble estimator defined as $f_{ens}(., \theta_{1:M}) = \frac{1}{M} \sum_{m=1}^{M} f(., \theta_m)$ where $\theta_{1:M} = \{\theta_1, \dots, \theta_M\}$ and each $\theta_m$or equivalently $f(., \theta_m)$ is a sample from the learning procedure. Let us denote $e_{1:M} = \{e_1, \dots, e_M\}$, so that the expectation over the distribution of the training procedures $e_1 \dots e_M$ writes as $E_{e_{1:M}}[.] = E_{e_M}\left[E_{e_{M-1}}\left[\dots E_{e_1}[.]\right]\right]$.

Let us define the bias and variance for f_{ens} :

$$bias(f_{ens}|(x, y)) = y - E_{e\ 1:M}\left[\frac{1}{M} \sum_{m=1}^{M} f(x, \theta_m)\right]$$

$$var(f_{ens}|x) = E_{e_{1:M}}[\left(\frac{1}{M}\sum_{m=1}^{M} f(x,\theta_m) - E_{e_{1:M}}\left[\frac{1}{M}\sum_{m=1}^{M} f(x,\theta_m)\right]\right)^2]$$

2.1 Preliminaries, show:

$$E_{e_{1:M}}\left[\frac{1}{M}\sum_{m=1}^{M} f(x,\theta_m)\right] = \frac{1}{M}\sum_{m=1}^{M} E_{e_m}[f(x,\theta_m)]$$

$$(\sum_{m=1}^{M} a_m - b_m)^2 = \sum_{m=1}^{M}(a_m - b_m)^2 + \sum_{m=1}^{M}\sum_{m'\neq m}(a_m - b_m)\,(a_{m'} - b_{m'})$$

Then using (1) for f_{ens}, show:

$$\mathcal{L}_{e_{1:M}}(f_{ens}) \triangleq E_{e_{1:M}}[\mathcal{L}(f_{ens},\theta_{1:M})] = E_{(x,y)\sim p}[B^2 + \frac{1}{M}V + \frac{M-1}{M}C]$$

Where (for simplification in the following $f_m(.)$ denotes $f(.,\theta_m)$):

$$B = \frac{1}{M}\sum_{m=1}^{M} bias\big(f_m|(x,y)\big)\,,\; V = \frac{1}{M}\sum_{m=1}^{M} var(f_m|x)\,,$$

$$C = \frac{1}{M(M-1)}\sum_m \sum_{m'\neq m} cov(f_m, f_{m'}|x) \text{ with } cov(f,f'|x) = E_{e,e'}[(f - E_e[f])(f' - E_{e'}[f'])]$$

2.2 Let us now suppose that the f_m are identically distributed. This means that the bias, variance and expectation are the same for all the f_m. Let us denote respectively $bias\big(f|(x,y)\big), var(f|x), \bar{f}$ these different variables – defined as above (question 1). Show that:

$$E_{e_{1:M}}[\mathcal{L}(f_{ens},\theta_{1:M})] = E_{(x,y)\sim p}[bias^2\big(f|(x,y)\big) + \frac{1}{M}var(f|x) + \frac{M-1}{M}cov(f,f'|x)] \qquad (2)$$

Where $cov(f,f'|x) = E_{e,e'}[\Big(f(x,\theta(e)) - \bar{f}(x)\Big)\Big(f(x,\theta(e')) - \bar{f}(x)\Big)]$.

2.3 Interpret this last result. What should be the property of the f functions for minimizing this generalization error?

2.4 We will now suppose that the f_m are independent, meaning that $cov(f,f'|x) = 0$, what is the expression of $E_{e_{1:M}}[\mathcal{L}(f_{ens},\theta_{1:M})]$? What is the benefit of using an ensemble method compared to a single function estimator f?

3. We will now consider an alternative to building ensembles, that emerged with the use of large pre-trained networks. We start from a pre-trained network (for example on ImageNet). Then we fine tune it on a given dataset using some learning procedure e. For different but related learning procedures e, one will get different set of parameters $\theta(e)$ that will be close one to the other. Suppose we fine tune M functions $f_m, m = 1\ldots, M$ Let us define the ensemble function:

$$f_{wa} \triangleq f(.,\theta_{wa}) \text{ with } \theta_{wa} = \frac{1}{M}\sum_{m=1}^{M}\theta_m$$

This means that the ensemble is defined by a unique function f_{wa}, its weights being the average of the weights of the M functions f_m. θ_{wa} is supposed close to $\theta_m\,, \forall m$. The f_m are supposed identically distributed so that (2) applies.

3.1 Using a first order Taylor expansion of $f(.,\theta_m)$ around $f(.,\theta_{wa})$, show that $f_{ens} - f_{wa} = O(\Delta_{1:M}^2)$ with $\Delta_{1:M} = \max_m \|\theta_m - \theta_{wa}\|_2$

3.2 Using a zeroth order Taylor expansion w.r.t. its first argument of $l(f_{ens}(x),y)$ around $f_{wa}(x)$, show that $l(f_{ens}(x),y) = l(f_{wa}(x),y) + O(\Delta_{1:M}^2)$

3.3 Using this result show that $\mathcal{L}(f_{wa},\theta_{1:M}) = \mathcal{L}(f_{ens},\theta_{1:M}) + O(\Delta_{1:M}^2)$

Show then $E_{e_{1:M}}[\mathcal{L}(f_{wa},\theta_{1:M})] = E_{e_{1:M}}[\mathcal{L}(f_{ens},\theta_{1:M})] + O(\bar{\Delta}^2)$ with $\bar{\Delta}^2 = E_{e_{1:M}}[\Delta_{1:M}^2]$

3.4 Interpret this result. What could be the benefit of the f_{wa} estimator compared to the f_{ens} estimator?

Note - Taylor expansion. Let $f: \mathbb{R}^p \to \mathbb{R}$ a differentiable function, the Taylor expansion of order 1 around point a is: $f(a+h) = f(a) + \nabla f(a).h + O(||h||^2)$