Introduction to Data Science - 1MS041

Benny Avelin

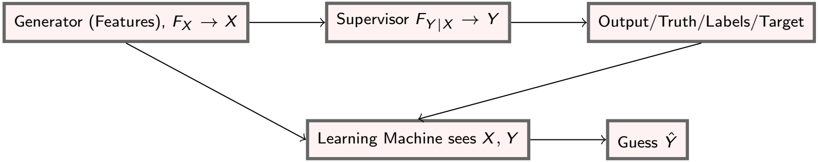
Department of Mathematics

HT 2023

Supervised learning

Setup

1. The generator of the data G
2. The supervisor S
3. The learning machine LM.



Pattern recognition

- Supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \ldots, k-1$.

Pattern recognition

- Supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \ldots, k-1$.
- Model space $\mathcal{M} = \{g_\lambda(x) : g_\lambda(x) \in \{0, \ldots, k-1\}\}$.

Pattern recognition

- Supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \dots, k-1$.
- Model space $\mathcal{M} = \{g_\lambda(x) : g_\lambda(x) \in \{0, \dots, k-1\}\}$.
- g_λ a **decision function**, **decision rule**, **classifier**.

Pattern recognition

- Supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \ldots, k-1$.
- Model space $\mathcal{M} = \{g_\lambda(x) : g_\lambda(x) \in \{0, \ldots, k-1\}\}$.
- g_λ a **decision function**, **decision rule**, **classifier**.
- $0 - 1$ loss

$$L(z, u) = \begin{cases} 0 & \text{if } y = u \\ 1 & \text{if } y \neq u \end{cases}$$

Pattern recognition

- Supervisors conditional distribution $F(y|x)$ is discrete, and can take k different values, $y = 0, \dots, k-1$.
- Model space $\mathcal{M} = \{g_\lambda(x) : g_\lambda(x) \in \{0, \dots, k-1\}\}$.
- g_λ a **decision function**, **decision rule**, **classifier**.
- $0 - 1$ loss

$$L(z, u) = \begin{cases} 0 & \text{if } y = u \\ 1 & \text{if } y \neq u \end{cases}$$

The pattern recognition problem

Minimize

$$R(\lambda) = \int L(y, g_\lambda(x)) dF(x, y) = \mathbb{E}[L(Y, g_\lambda(X))]$$

where $(X, Y) \sim F(x, y)$, where $g_\lambda \in \mathcal{M}$.

Recall that,

$$\mathbb{E}[L(Y, g_\lambda(X))] = \mathbb{P}(\{Y \neq g_\lambda(X)\}).$$

We want to minimize the empirical version of the risk.

Recall that,

$$\mathbb{E}[L(Y, g_\lambda(X))] = \mathbb{P}(\{Y \neq g_\lambda(X)\}).$$

We want to minimize the empirical version of the risk.

Definition

Assume that $Z = ((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)) \overset{\text{IID}}{\sim} F(x, y)$ is a sequence of $\mathbb{R}^{m+1}$ valued random variables taking values in the data space $\mathbb{X} \times \mathbb{Y}$. We define the empirical risk for a function $g : \mathbb{X} \to \mathbb{Y}$ as

$$\hat{R}_n(g) = \hat{R}_n(Z; g) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, g(X_i)).$$

Recall that,

$$\mathbb{E}[L(Y, g_\lambda(X))] = \mathbb{P}(\{Y \neq g_\lambda(X)\}).$$

We want to minimize the empirical version of the risk.

Definition

Assume that $Z = ((X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)) \overset{\text{IID}}{\sim} F(x, y)$ is a sequence of $\mathbb{R}^{m+1}$ valued random variables taking values in the data space $\mathbb{X} \times \mathbb{Y}$. We define the empirical risk for a function $g : \mathbb{X} \to \mathbb{Y}$ as

$$\hat{R}_n(g) = \hat{R}_n(Z; g) = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, g(X_i)).$$

Given a model space $\mathcal{M}$ we consider

$$\hat{g}_n^* := \hat{g}_n^*(Z) := \arg\min_{g \in \mathcal{M}} \hat{R}_n(g).$$

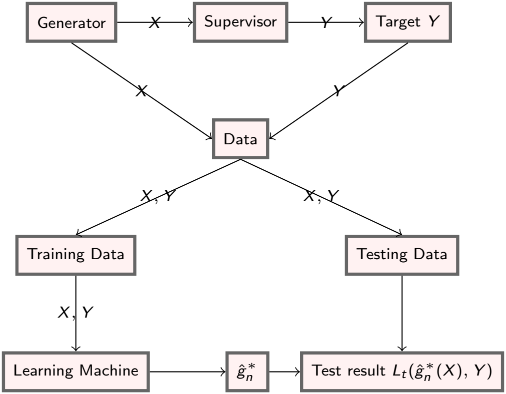
Trained model

The decision function

$$\hat{g}_n^* := \hat{g}_n^*(Z) := \arg\min_{g \in \mathcal{M}} \hat{R}_n(g)$$

is what we call the trained model.

Supervised learning



Test result

- Dataset $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$.

Test result

- Dataset $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$.
- $T_{rain} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ **Training data**

Test result

- Dataset $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$.
- $T_{rain} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ **Training data**
- $T_{est} = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$ **Test data**

Test result

- Dataset $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$.
- $T_{rain} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ **Training data**
- $T_{est} = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$ **Test data**
- The trained model

$$\hat{g} = \arg\min_{\phi \in \mathcal{M}} \hat{R}_n(g)$$

Test result

- Dataset $T_{n+m} := \{(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})\}$ sampled i.i.d from $F_{X,Y}$.
- $T_{rain} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ **Training data**
- $T_{est} = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$ **Test data**
- The trained model

$$\hat{g} = \arg\min_{\phi \in \mathcal{M}} \hat{R}_n(g)$$

- The empirical testing error is

$$\hat{R}_m(\hat{g}) = \frac{1}{m} \sum_{i=n+1}^{m} L(\hat{g}(X_i), Y_i)$$

Testing

Testing

When testing, we are considering the training data as given and the model $\hat{g}$ as fitted and fixed. We then test this model on the unseen data (Testing data) and evaluate the loss on the testing data. This is our testing error!

Testing

Testing

When testing, we are considering the training data as given and the model $\hat{g}$ as fitted and fixed. We then test this model on the unseen data (Testing data) and evaluate the loss on the testing data. This is our testing error!

Our true testing error is

$$\mathbb{E}[L(\hat{g}(X), Y) \mid T_{rain}]$$

that is, given the training data T_{rain} we want the empirical test error to be an estimate of the true test error.

Point of view

Two points of views

- Make a lot of assumptions about the data, and make guarantees before we fit our model. Think linear regression assuming everything is Gaussian.

Point of view

Two points of views

- Make a lot of assumptions about the data, and make guarantees before we fit our model. Think linear regression assuming everything is Gaussian.
- We make very little assumptions on data, fit our model and then test it. We want to design a test measurement that is well behaved no matter what (think bounded).

Main point in the train test framework

Main point

How do we think about this?

- We think of the actual fitting procedure on the training data as a black box.

Main point in the train test framework

Main point

How do we think about this?

- We think of the actual fitting procedure on the training data as a black box.
- This black box gives out a decision function, we say, OK lets test it.

Main point in the train test framework

Main point

How do we think about this?

- We think of the actual fitting procedure on the training data as a black box.
- This black box gives out a decision function, we say, OK lets test it.
- We perform a test on data that never went into the black box and want to give guarantees for the empirical test error!

Main point in the train test framework

Main point

How do we think about this?

- We think of the actual fitting procedure on the training data as a black box.
- This black box gives out a decision function, we say, OK lets test it.
- We perform a test on data that never went into the black box and want to give guarantees for the empirical test error!
- Thus we can just choose a different measurement than the loss.

Metrics

- We have seen the Risk as a way to measure a decision function, this is a performance metric.

Metrics

- We have seen the Risk as a way to measure a decision function, this is a performance metric.
- We are free to choose the underlying loss function L when testing.

Metrics

- We have seen the Risk as a way to measure a decision function, this is a performance metric.
- We are free to choose the underlying loss function L when testing.
- We can also consider conditional Risk.

Metrics

1. Consider the conditional risk

$$R_1(\lambda) = \mathbb{P}(Y = 1 \mid g_\lambda(X) = 1) = \mathbb{E}[\mathbb{1}_{Y=1} \mid g_\lambda(X) = 1]$$

this goes by the name **precision**.

Metrics

1. Consider the conditional risk

$$R_1(\lambda) = \mathbb{P}(Y = 1 \mid g_\lambda(X) = 1) = \mathbb{E}[\mathbb{1}_{Y=1} \mid g_\lambda(X) = 1]$$

this goes by the name **precision**. The empirical version is

$$\frac{\sum_{i=1}^m \mathbb{1}_{Y_i=1} \mathbb{1}_{g_\lambda(X_i)=1}}{\sum_{i=1}^m \mathbb{1}_{g_\lambda(X_i)=1}} = \frac{\text{nr of correctly prediced 1}}{\text{nr of predicted 1}}$$

Metrics

1. Consider the conditional risk

$$R_1(\lambda) = \mathbb{P}(Y = 1 \mid g_\lambda(X) = 1) = \mathbb{E}[\mathbb{1}_{Y=1} \mid g_\lambda(X) = 1]$$

this goes by the name **precision**. The empirical version is

$$\frac{\sum_{i=1}^m \mathbb{1}_{Y_i=1}\mathbb{1}_{g_\lambda(X_i)=1}}{\sum_{i=1}^m \mathbb{1}_{g_\lambda(X_i)=1}} = \frac{\text{nr of correctly prediced 1}}{\text{nr of predicted 1}}$$

2. The conditional risk

$$R_2(\lambda) = \mathbb{P}(g_\lambda(X) = 1 \mid Y = 1) = \mathbb{E}[\mathbb{1}_{g_\lambda(X)=1} \mid Y = 1]$$

this goes by the name **recall**.

Metrics

1. Consider the conditional risk

$$R_1(\lambda) = \mathbb{P}(Y = 1 \mid g_\lambda(X) = 1) = \mathbb{E}[\mathbb{1}_{Y=1} \mid g_\lambda(X) = 1]$$

   this goes by the name **precision**. The empirical version is

$$\frac{\sum_{i=1}^m \mathbb{1}_{Y_i=1}\mathbb{1}_{g_\lambda(X_i)=1}}{\sum_{i=1}^m \mathbb{1}_{g_\lambda(X_i)=1}} = \frac{\text{nr of correctly prediced 1}}{\text{nr of predicted 1}}$$

2. The conditional risk

$$R_2(\lambda) = \mathbb{P}(g_\lambda(X) = 1 \mid Y = 1) = \mathbb{E}[\mathbb{1}_{g_\lambda(X)=1} \mid Y = 1]$$

   this goes by the name **recall**. The empirical version is

$$\frac{\sum_{i=1}^m \mathbb{1}_{Y_i=1}\mathbb{1}_{g_\lambda(X_i)=1}}{\sum_{i=1}^m \mathbb{1}_{g_\lambda(X_i)=1}} = \frac{\text{nr of correctly prediced 1}}{\text{nr of actual 1}}$$

Guarantees

Setup

Consider now the $0-1$ loss L, and let us be given a decision function g. Let us also assume that we have some i.i.d. testing data $T_{est} = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$.

Guarantees

Setup

Consider now the $0-1$ loss L, and let us be given a decision function g. Let us also assume that we have some i.i.d. testing data $T_{est} = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$.

- $L(g(X_i), Y_i) \in \{0, 1\}$ is certainly a bounded random variable.

Guarantees

Setup

Consider now the $0-1$ loss L, and let us be given a decision function g. Let us also assume that we have some i.i.d. testing data $T_{est} = \{(X_1, Y_1), \dots, (X_m, Y_m)\}$.

- $L(g(X_i), Y_i) \in \{0, 1\}$ is certainly a bounded random variable.
- Since $\{(X_1, Y_1), \dots, (X_m, Y_m)\}$ is an i.i.d. seqence, so is $L(g(X_1), Y_1), \dots, L(g(X_m), Y_m)$.

Guarantees

Setup

Consider now the $0-1$ loss L, and let us be given a decision function g. Let us also assume that we have some i.i.d. testing data $T_{est} = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$.

- $L(g(X_i), Y_i) \in \{0, 1\}$ is certainly a bounded random variable.
- Since $\{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ is an i.i.d. seqence, so is $L(g(X_1), Y_1), \ldots, L(g(X_m), Y_m)$.
- Denote $Z_i = L(g(X_i), Y_i)$, by using Hoeffdings inequality we thus get

$$\mathbb{P}(|\overline{Z}_m - \mathbb{E}[Z_1]| \geq \epsilon) \leq 2e^{-2m\epsilon^2}.$$

Confidence intervals

Then for $\alpha \in (0, 1)$ we have for $\delta = \frac{1}{\sqrt{m}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$

$$\mathbb{P}(\overline{Z}_m - \delta \leq \mathbb{E}[Z_1] \leq \overline{Z}_m + \delta) \geq 1 - \alpha.$$

Confidence intervals

Then for $\alpha \in (0,1)$ we have for $\delta = \frac{1}{\sqrt{m}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$

$$\mathbb{P}(\overline{Z}_m - \delta \leq \mathbb{E}[Z_1] \leq \overline{Z}_m + \delta) \geq 1 - \alpha.$$

- Recall that $\mathbb{E}[Z_1] = \mathbb{E}[L(g(X_1), Y_i)]$ i.e. the true risk.

Confidence intervals

Then for $\alpha \in (0,1)$ we have for $\delta = \frac{1}{\sqrt{m}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$

$$\mathbb{P}(\overline{Z}_m - \delta \leq \mathbb{E}[Z_1] \leq \overline{Z}_m + \delta) \geq 1 - \alpha.$$

- Recall that $\mathbb{E}[Z_1] = \mathbb{E}[L(g(X_1), Y_i)]$ i.e. the true risk.
- $\overline{Z}_n$ is thus our empirical risk.

Confidence intervals for Metrics

- In order to use something like Hoeffdings inequality we need to write the empirical metric as a sum. So we should rewrite our conditional metrics in terms of conditional random variables.

Confidence intervals for Metrics

- In order to use something like Hoeffdings inequality we need to write the empirical metric as a sum. So we should rewrite our conditional metrics in terms of conditional random variables.
- Consider again

$$R_1(\lambda) = \mathbb{P}(Y = 1 \mid g_\lambda(X) = 1) = \mathbb{E}[\mathbb{1}_{Y=1} \mid g_\lambda(X) = 1]$$

Define the conditional random variable $Z = Y \mid (g_\lambda(X) = 1)$, then we will get $Z_1, \ldots, Z_k$ coming from $(X_1, Y_1), \ldots, (X_m, Y_m)$ where k is the number of observations for which $g_\lambda(X_i) = 1$.

Confidence intervals for Metrics

- In order to use something like Hoeffdings inequality we need to write the empirical metric as a sum. So we should rewrite our conditional metrics in terms of conditional random variables.
- Consider again

$$R_1(\lambda) = \mathbb{P}(Y = 1 \mid g_\lambda(X) = 1) = \mathbb{E}[\mathbb{1}_{Y=1} \mid g_\lambda(X) = 1]$$

  Define the conditional random variable $Z = Y \mid (g_\lambda(X) = 1)$, then we will get $Z_1, \ldots, Z_k$ coming from $(X_1, Y_1), \ldots, (X_m, Y_m)$ where k is the number of observations for which $g_\lambda(X_i) = 1$.
- $Z_i \in \{0, 1\}$ so we can again use Hoeffding, but note, for $\alpha \in (0, 1)$ we have for $\delta = \frac{1}{\sqrt{k}}\sqrt{\frac{1}{2}\ln\left(\frac{2}{\alpha}\right)}$

$$\mathbb{P}(\overline{Z}_k - \delta \leq \mathbb{E}[Z_1] \leq \overline{Z}_k + \delta) \geq 1 - \alpha.$$