Introduction to Data Science - 1MS041

Benny Avelin
Department of Mathematics

HT 2023

Recap

Different problems

Find f

In the problem "find f", the supervisor uses f to construct Y, i.e. $Y = f(X)$ and the learning machine tries to find f.

Different problems

Find f

In the problem "find f", the supervisor uses f to construct Y, i.e. $Y = f(X)$ and the learning machine tries to find f.

Regression

In the regression problem, the supervisor has $F_{Y|X}$ but the Learning machine only tries to learn $r(X) = \mathbb{E}[Y \mid X]$.

Different problems

Find f

In the problem "find f", the supervisor uses f to construct Y, i.e. $Y = f(X)$ and the learning machine tries to find f.

Regression

In the regression problem, the supervisor has $F_{Y|X}$ but the Learning machine only tries to learn $r(X) = \mathbb{E}[Y \mid X]$.

Pattern recognition

The supervisor uses a discrete distribution for Y, i.e. $F_{Y|X}$ is a discrete distribution. Here the learning machine uses $0 - 1$ loss.

Regression

Regression function

$$r(x) = \int y dF(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

Regression

Regression function

$$r(x) = \int y dF(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

We need some things to be defined

1. Statistical model:
   $\mathcal{F} = \{F(x, y) = \mathbb{1}_{y \geq f(x)}(y) F(x), f \in C([0,1], \mathbb{R})\}$

Regression

Regression function

$$r(x) = \int y dF(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

We need some things to be defined

1. Statistical model:
   $\mathcal{F} = \{F(x,y) = \mathbb{1}_{y \geq f(x)}(y) F(x), f \in C([0,1], \mathbb{R})\}$
2. Model space: $\mathcal{M} = \{g_\lambda(x) : \lambda \in \Lambda\}$, a parametrized space of functions in which we are searching for f, or an approximation thereof. (The kind of functions the learning machine can represent).

Regression

Regression function

$$r(x) = \int y dF(y|x) = \mathbb{E}[Y \mid X = x].$$

Here, r is called the regression function.

We need some things to be defined

1. Statistical model:
   $\mathcal{F} = \{F(x, y) = \mathbb{1}_{y \geq f(x)}(y) F(x), f \in C([0,1], \mathbb{R})\}$
2. Model space: $\mathcal{M} = \{g_\lambda(x) : \lambda \in \Lambda\}$, a parametrized space of functions in which we are searching for f, or an approximation thereof. (The kind of functions the learning machine can represent).
3. A loss function $L : [0,1] \times \mathbb{R} \to \mathbb{R}$.

Regression

Consider the quadratic risk

$$R(\phi) = \mathbb{E}[(Y - g(X))^2]$$

Then

$$R(\lambda) = \mathbb{E}[(Y - g_\lambda(X))^2] = \mathbb{E}[(Y - r(X))^2] + \mathbb{E}[(r(X) - g_\lambda(X))^2]$$

Here the first term is the variance and the second term is bias^2.

How do we measure a regression model?

Train-test

We again use the Train-test framework. That is, we train our model $\hat{\phi}$ on the training data and test it on the test data.

Common test metrics

- The mean squared error (MSE) (Usually used to measure model fit)

$$\mathbb{E}[(\hat{\phi}(X) - Y)^2 \mid \hat{\phi}]$$

and the root mean squared error (RMSE)

$$\sqrt{\mathbb{E}[(\hat{\phi}(X) - Y)^2 \mid \hat{\phi}]}.$$

- The mean absolute error (MAE)

$$\mathbb{E}[|\hat{\phi}(X) - Y| \mid \hat{\phi}]$$

often preferred as it is more explanatory.

- R^2, or explained variance

$$1 - \frac{\mathbb{E}[(\hat{\phi}(X) - Y)^2 \mid \hat{\phi}]}{\mathbb{V}(Y)}$$

Calibration

Calibration error

Consider f a given fixed function, then the calibration error is defined as

$$\sqrt{\mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|^2]}$$

Note that

$$\mathbb{E}[\|Y - f(X)\|^2] = \mathbb{E}[\|\mathbb{E}[Y \mid f(X)] - f(X)\|^2] + \mathbb{E}[\|Y - \mathbb{E}[Y \mid f(X)]\|^2]$$

here we think about the first term as the bias^2 and the second term as variance. Thus we should interpret the calibration error as bias. The variance term should be considered as the variance of the prediction.

Calibration

Note

The calibration of models has become very important, and is an active area of research. Problem is, it is not super easy to compute as it is essentially another regression problem, i.e. to compute $\mathbb{E}[Y \mid f(X)]$ is a regression problem.

Confidence intervals of the metrics

Assumptions

- For MSE, and MAE we have to assume that the quantity $(Y - f(X))^2$ is sub-Gaussian or $|Y - f(X)|$, this can be hard to justify unless we for instance know that things are bounded.
- For R^2, we have to produce a confidence interval for a ratio, which is quite delicate, especially if the variance of Y is small. In order to get anything that makes sense we need to know something about the fourth moments.