STATISTICAL DECISION THEORY

23/02/2021

SUPERVISED LEARNING : Given observations (x_i, y_i), $i = 1, \dots, n$

find a rule $\hat{f}$ <
- regression function (quantitative response Y)
- classifier (qualitative response Y)

that allows to predict Y from X.

So $\hat{f}(X)$ will be some "approximation" of a true $f(X)$, describing the "true" relationship between Y and X.

What "true" is though depends on the way we set up our statistical learning problem (how would we like to describe/capture this relationship?)

Statistically, this boils down to choosing a <u>loss function</u> that decides in which way $f(X)$ should describe the data.

We are going to explore this in the regression and classification setting.

REGRESSION SETTING

In regression, a popular choice is the squared error loss:

SQUARED ERROR LOSS

$$L(Y, f(X)) = (Y - f(X))^2$$

Leading to the criterion: find f that minimizes

EXPECTED PREDICTION ERROR

$$EPE(f) = E_{X,Y}\left[(Y - f(X))^2\right]$$

We will see that setting the problem in this way is equivalent to considering $f(x) = E(Y \mid X = x)$, the conditional mean of Y given x, also called regression function.

Theorem $E[(Y-f(X))^2]$ has a minimum when

$$f(x) = E[Y \mid X=x] \quad \forall x$$

Proof

$$E[(Y-f(X))^2] = \iint (y-f(x))^2 g_{X,Y}(x,y)\,dx\,dy$$

($g_{X,Y}$: joint density of X and Y)

$$= \iint (y-f(x))^2 g_{Y|X}(y|x)\, g_X(x)\,dy\,dx$$

$$= \int \left[\int (y-f(x))^2 g_{Y|X}(y|x)\,dy\right] g_X(x)\,dx$$

$$\left[\int (y-f(x))^2 g_{Y|X}(y|x)\,dy\right] = E_{Y|X}[(Y-f(X))^2 \mid X=x]$$

$g_{X,Y}(x,y) = g_{Y|X}(y|x)\, g_X(x)$

Iterated rule of expectation:

$E_{X,Y}(h(X,Y)) = E_X[E_{Y|X}[h(X,Y)|X]]$

Note: $Z \sim h(Z)$

$\min_a E[(Z-a)^2] = E[(Z-E(Z))^2]$

$\frac{d}{da} E(Z^2 - 2aZ + a^2) =$

$= \frac{d}{da}(E(Z^2) - 2a E(Z) + a^2) = -2E(Z) + 2a$

$\Rightarrow \hat{a} = E(Z)$!

Therefore $\int (y-f(x))^2 g_{Y|X}(y|x)\,dy \geq \int (y-E(Y|X=x))^2 g_{Y|X}(y|x)\,dy$.

But this is true for all x, so $E[(Y-f(X))^2] \geq E[(Y-E(Y|X))^2]$ ∎

SQUARED ERROR LOSS $\longleftrightarrow$ $f(x) = E(Y|X=x)$

① Different losses would lead to different functions,

eg $L(Y, f(X)) = |Y - f(X)| \rightarrow f(x) = \text{median}(Y|X=x)$ QUANTILE REGRESSION

② Statistical learning under a squared error loss consists in different methods for estimating $E(Y|X=x)$, which is unknown.

For example:

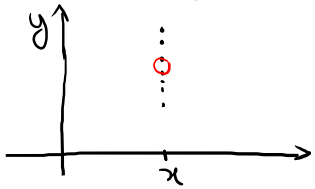
- LINEAR REGRESSION (parametric)

$$E(Y|X=x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

with $\beta_0, \beta_1, \ldots, \beta_p$ estimated from training data (least squares).

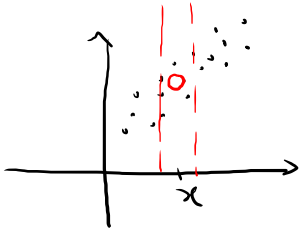
- **K-NEAREST NEIGHBOUR (non-parametric)**

At each point x, calculate the average (sample mean) of all training points y_i with $x_i = x$.



Typically only few observations for each x, often only one, so

$$\hat{f}(x) = \text{Average}\left(y_i \mid x_i \in \underbrace{N_K(x)}_{\substack{\text{neighbourhood:} \\ K \text{ points closest to } x}} \right)$$



Two approximations:

1. Expectation approximated by sample mean
2. Conditioning on a point relaxed by conditioning on a region around that point.

Generally good only for very large sample sizes (hence small neighbourhood) and also small p. This simple method however inspired more complex methods, called Kernel methods (which use weights that decrease smoothly as we go away from the target point).

CLASSIFICATION SETTING

What is a reasonable loss function here?

Response Y is categorical, with the typical case BINARY (yes/no).

A classifier is therefore going to decide which class to assign an object x.

Our $\hat{Y}$ in this case is therefore a predicted class and the "error" that we commit is to assign x to the wrong class.

Since predictions are classes, we can represent the loss as a $K \times K$ table, with K the number of classes.

In the binary case:

ZERO-ONE LOSS

$$L(Y, f(X)) =$$

TRUE \ PREDICTED	0	1
0	0	1
1	1	0

This leads to the criterion: find $f(X)$ that minimizes

EXPECTED 0-1 LOSS

$$E_{X,Y}\, L(Y, f(X))$$

MINIMIZING EXPECTED 0-1 LOSS

Let us consider a fixed x. If we predict x to class 0 ($\hat{y}=0$) then

$$E[L(Y,0)] = 0 \cdot p(0|x) + 1 \cdot p(1|x) = p(1|x)$$

Whereas, if we predict x to class 1 ($\hat{y}=1$),

$$E[L(Y,1)] = 0 \cdot p(1|x) + 1 \cdot p(0|x) = p(0|x)$$

So according to 0-1 loss, we should classify x to class 1 if

$$p(1|x) > p(0|x), \text{ and } 0 \text{ otherwise.}$$

In general, with more than 2 classes, the loss is minimized by assigning each observation to the most likely class given the predictor values:

$$\text{ASSIGN } x \text{ to CLASS } j = \underset{c \in \text{classes}}{\operatorname{argmax}} P(Y=c|x)$$

BAYES CLASSIFIER

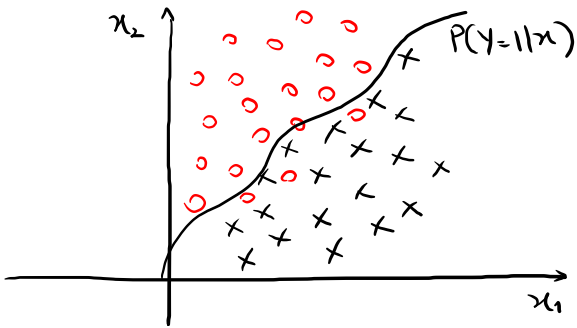
BAYES CLASSIFIER

Going back to the two classes, the Bayes classifier results in:

ASSIGN x to CLASS $\begin{cases} 1 & \text{if } P(1|x) > 0.5 \\ 0 & \text{if } P(1|x) < 0.5 \end{cases}$

$P(1|x) > P(0|x)$
$\Rightarrow P(1|x) > 1 - P(1|x)$
$\Rightarrow 2P(1|x) > 1$
$\Rightarrow P(1|x) > 0.5$

BAYES DECISION BOUNDARY : All $\underline{x}$ such that $P(Y=1|\underline{x}) = 0.5$



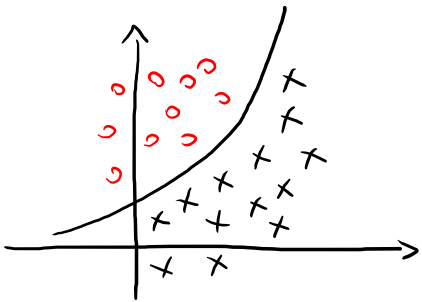
O = Class 1
X = Class 0

BAYES ERROR RATE

The overall error that one commits can be defined by

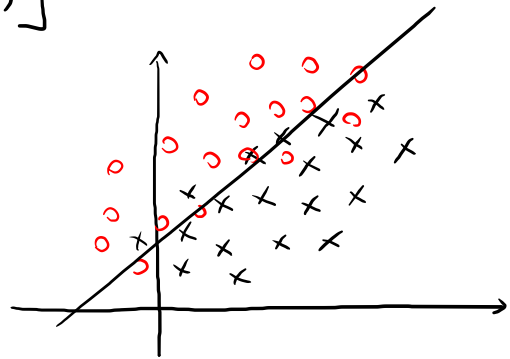
BAYES ERROR RATE

$$BER = 1 - E_x\left[\max_c P(y=c|x)\right]$$



NO OVERLAP, BER = 0

(For each x, $\max_c P(y=c|x)$ is 1 for one class and zero for all the others)



OVERLAP, BER > 0

(most cases, analogous to irreducible error in regression problems)

MORE GENERAL LOSS FUNCTION

A 0-1 Loss assumes equal costs of misclassification:

0	1
1	0

This can be relaxed by choosing a more general loss function:

MISCLASSIFICATION LOSS

TRUE \ PREDICTED	0	1
0	0	C_0
1	C_1	0

Useful in many contexts:

1. One misclassification more costly than another (wrong medical diagnosis, giving mortgages to bad customers, ...)
2. Unbalanced class sizes (only 1% in one class so 99% if we classify everything to the large class)

In this case:

$$E[L(Y,0)] = C_1 p(1|x)$$

$$E[L(Y,1)] = C_0 p(0|x)$$

$\Rightarrow$ ASSIGN x to class 1 if

$$C_1 p(1|x) > C_0 p(0|x)$$

$p(0|x) = 1 - p(1|x)$

$$\Rightarrow \quad p(1|x) > \frac{C_0}{C_1 + C_0}$$

NEW THRESHOLD DEFINING THE DECISION BOUNDARY (0.5 if $C_0 = C_1$)

REGRESSION vs CLASSIFICATION SETTING

REGRESSION Under a squared error loss, we are interested in

$$f(x) = E(Y \mid X = x)$$

(Different methods will provide different estimates.)

CLASSIFICATION (binary) Assign x to the class 1 according to

$$f(x) = P(Y = 1 \mid X = x)$$

($P(Y = 1 \mid X = x)$ is unknown. Different methods will provide different ways of estimating this quantity.)

ASSESSING MODEL ACCURACY

In practice, we use our chosen method to estimate $f(x)$ from a training data (x_i, y_i), $i = 1, \ldots, n$ and then test its performance on some test data $(x_i^{(t)}, y_i^{(t)})$, $i = 1, \ldots, m$ (unseen data, not used to build $\hat{f}(x)$).

REGRESSION

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \left(y_i^{(t)} - \hat{f}(x_i^{(t)})\right)^2$$

MEAN-SQUARED ERROR

CLASSIFICATION

(Counting misclassification errors)

(Codes: 0, 1, binary)

CONFUSION MATRIX

TRUE \ PREDICTED	0	1
0	TN	FP
1	FN	TP

TN: True Negative

TP: True Positive

FP: False Positive

FN: False Negative

$m = TN + TP + FP + FN$

ERROR RATE (ON TEST DATA)

TRUE \ PREDICTED	0	1
0	TN	FP
1	FN	TP

$\Rightarrow$

$$TPR = \frac{TP}{TP+FN} \qquad \text{SENSITIVITY}$$

$$FPR = \frac{FP}{TN+FP} = 1 - \frac{TN}{TN+FP} \qquad 1 - \text{SPECIFICITY}$$

ERROR RATE

$$\frac{FN+FP}{TN+TP+FN+FP} = \frac{FN+FP}{m} = \frac{1}{m}\sum_{i=1}^{m}\left(y_i^{(i)} \neq \hat{y}_i^{(i)}\right)$$

If $C_0 \neq C_1$:

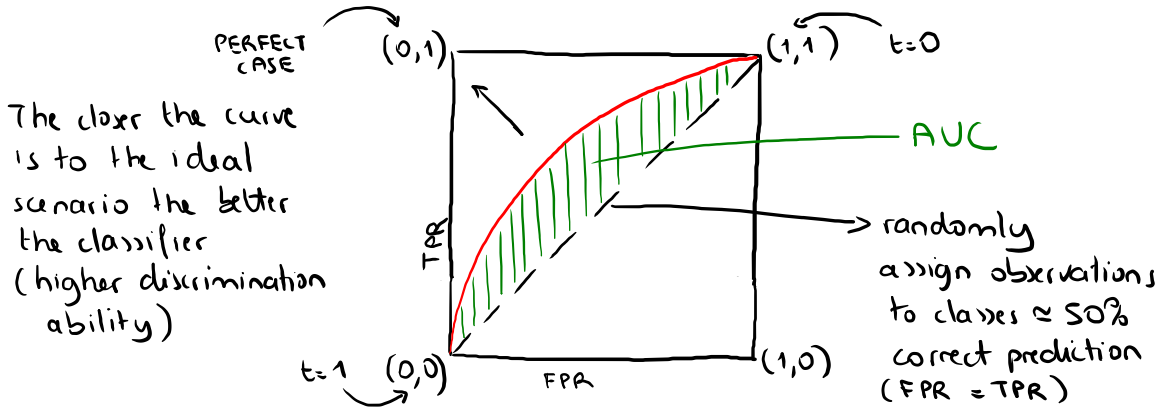
MISCLASSIFICATION COST

$$: \quad \frac{C_0 FP + C_1 FN}{m} \qquad \text{(weighted error rate)}$$

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE

Since misclassification costs may not be known exactly, it is also common to evaluate a classifier across the full range of thresholds

ROC curve : Plots TP vs FP for $t \in [0,1]$



The closer the curve is to the ideal scenario the better the classifier (higher discrimination ability)

It is common to summarise the curve with the Area Under the Curve (AUC) as a single measure of accuracy. Different models can be compared based on their AUCs, though careful with crossing curves!