Threshold classifier

- hard response, ~~...~~ $X \in \mathbb{R}$ (1-dimensional), $Y \in \{-1, +1\}$

decision rule
$$\hat{Y} = \begin{cases} 1 & \text{if } X > X_0 \\ -1 & \text{if } X < X_0 \end{cases} \quad (X = X_0) \text{ arbitrary}$$

X_0 ← threshold

- example: X is a person's weight $Y = 1$ obese $Y = -1$ normal
  $X_0 = 100$ kg, but this is not a very good method, because it doesn't adjust for a person's height

⇒ solution for several features: design a formula that combines the different features into a single number ($\hat{=}$ new feature)

- example: $X \in \mathbb{R}^2$ X_1: weight [kg] X_2: height [m] $Y = 1$ obese

body-mass-index $BMI = \dfrac{\text{weight}}{\text{height}^2}$

decision rule $Y = 1$ if $BMI > 30$

actually: use multiple thresholds for more informative labels

$$\hat{Y} = \begin{cases} \text{severe underweight} & \text{if } BMI < 16 \\ \text{underweight} & BMI < 18.5 \\ \text{normal} & < 25 \\ \text{over weight} & < 30 \\ \text{obese} & \geq 30 \end{cases}$$

threshold classifier: given $X \in \mathbb{R}$, predict label $\hat{Y} = \begin{cases} 1 & \text{if } X \geq t \\ 0 & \text{if } X < t \end{cases}$

[indicator function: $\mathbb{1}[\underbrace{\text{condition}(X)}_{\text{either true or false, depending on value of } X}] = \begin{cases} 1 & \text{if condition}(X) == \text{true} \\ 0 & \text{if condition}(X) == \text{false} \end{cases}$

in programming: type conversion from $\underset{\text{condition}}{\text{boolean}}$ to $\underset{1 \text{ or } 0}{\text{reals}}$

in C/C++ : `double result = (double) (x > t);`]

rewrite threshold: $\hat{Y} = \mathbb{1}[X \geq t]$

compare with Bayesian classifier (= best possible)

define a toy problem:
- simplified "cartoon" of some real world problem
- is not necessarily practically relevant
- but: we learn a lot about the machine learning method that we use

here: $X \in [0, 1]$ prior of Y: $p(Y=1) = p(Y=0) = 1/2$

[this notation actually is $p^*(Y)$, but we drop the $*$ for simplicity]

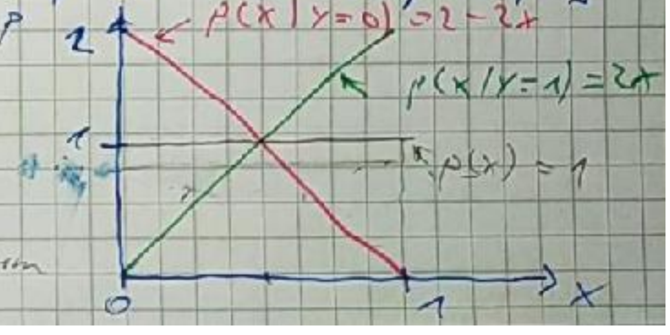
likelihoods

$p(X \mid Y=0) = 2 - 2x$

$p(X \mid Y=1) = 2x$

evidence:

$$p(x) = \sum_{k=0,1} p(x \mid Y=k)\, p(Y=k)$$
$$= (2-2x) \cdot \frac{1}{2} + 2x \cdot \frac{1}{2}$$
$$= 1 - x + x = \underline{\underline{1}} \text{ uniform}$$



posteriors according to Bayes: $$p(Y|X) = \frac{p(X|Y)\,p(Y)}{p(X)}$$

$$p(Y=0|X) = \frac{(2-2x)\cdot\frac{1}{2}}{1} = 1-x \qquad p(Y=1|X) = \frac{2x\cdot\frac{1}{2}}{1} = x$$

Bayes classifier: $$y^* = \arg\max_k p(Y=k|X)$$
$$= \begin{cases} 0 & \text{if } (1-x) \geq x \iff x \leq 1/2 \\ 1 & \text{if } (1-x) < x \iff x > 1/2 \end{cases}$$
$$= \underline{\mathbb{1}[x > 1/2]} \;\hat{=}\; \text{threshold classifier with } t = 1/2$$

compare with all possible threshold classifiers:

- type A: $\hat{Y} = \mathbb{1}[X > t]$ - type B: $\hat{Y} = \mathbb{1}[X < t]$

compute the probability of error $p(\text{error} \mid \underbrace{\text{type}, t}_{\text{two degrees of freedom}})$

$$p(\text{error}|\text{type},t) = \mathbb{E}_{\substack{X\sim p(X|Y)\\ Y\sim p(Y)}}\Big[p\big(\mathbb{1}[\underbrace{\text{condition}(X)}_{\substack{\mathbb{1}[X>t] \text{ if type}=A \\ \mathbb{1}[X<t] \text{ if type}=B}}] \neq Y \mid \text{type}, t\big)\Big]$$

since $p(Y=1)=p(Y=0)=\frac{1}{2}$ for classifier A $\Big\}$ $$= \mathbb{E}_x\big[p(Y=1|X)\,\mathbb{1}[\underset{\substack{\text{error, since} \\ \text{opposite of type A} \\ \text{behavior}}}{X<t}]\big] + \mathbb{E}_x\big[p(Y=0|X)\,\mathbb{1}[X>t]\big]$$

$$\mathbb{E}_X\left[p(Y=1|X)\,\mathbb{1}[X<t]\right] = \int_0^1 \underbrace{p(Y=1|X)\,\mathbb{1}[X<t]}_{\text{argument of } \mathbb{E}}\ \underbrace{p(X)}_{\text{from } \mathbb{E}}\ dX$$

false negative

$$= \int_0^t \underbrace{p(Y=1|X)}_{=X}\ \underbrace{p(X)}_{=1}\ dX = \int_0^t X \cdot 1 \ dX = \frac{X^2}{2}\Big|_0^t = \frac{t^2}{2}$$

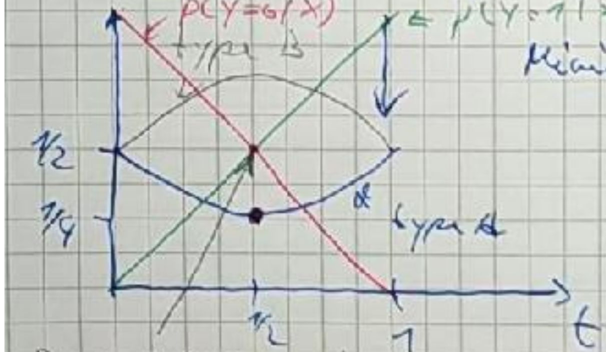
$$\mathbb{E}_X\left[p(Y=0|X)\,\mathbb{1}[X\geq t]\right] = \int_0^1 p(Y=0|X)\,\mathbb{1}[X\geq t]\ p(X)\ dX$$

false positive

$$= \int_t^1 \underbrace{p(Y=0|X)}_{=1-X}\ \underbrace{p(X)}_{=1}\ dX = \int_t^1 (1-X)\,dX = \left(X-\frac{X^2}{2}\right)\Big|_t^1$$

$$= 1 - \frac{1^2}{2} - \left(t - \frac{t^2}{2}\right) = \frac{1}{2} - t + \frac{t^2}{2} =$$

$$\mathbb{E}_X\left[p(\text{error}\,|\,\text{type}=A,\ t)\right] = \frac{t^2}{2} + \frac{1}{2} - t + \frac{t^2}{2} = t^2 - t + \frac{1}{2} = \left(t-\frac{1}{2}\right)^2 + \frac{1}{4}$$



Bayes error: set threshold where the curves cross

Minimum is achieved at $t = 1/2$, as in Bayes classifier

Same calculation for type B (reverse all conditions)

$$\mathbb{E}_X\left[p(\text{error}\,|\,\text{type}=B,\ t)\right] = -\left(t-\frac{1}{2}\right)^2 + \frac{3}{4}$$
$$= 1 - \mathbb{E}_X[\text{type } A]$$

Minimum is achieved for $t=0$ or $t=1$

$$\mathbb{E}_X[p(\text{error})] = 1/2 \ \hat{=} \text{ pure guessing}$$

in contrast, the best error for type A at $t = 1/2$ is $1/4$ [$\hat{=}$ Bayes rate] likelihoods are probability densities, because X is continuous

normalization $\int_0^1 p(X \mid Y)\, dx = 1$ for all Y [here: $Y=0$ or $Y=1$]

$$\int_0^1 p(x \mid Y=0)\, dx = \int_0^1 (2-2x)\, dx = 2x - \frac{\cancel{2}x^2}{\cancel{2}}\Big|_0^1 = 2-1-(0-0) = 1$$

$$\int_0^1 p(x \mid Y=1)\, dx = \int_0^1 2x\, dx = \cancel{2}\frac{x^2}{\cancel{2}}\Big|_0^1 = 1-0 = 1$$

How to generalize the threshold classifier when there are multiple features

$X \in \mathbb{R}^D$ => should be better, because more features $\hat{=}$ more information on Y

problem: comparison $X \geq t$ is only defined for scalars $X \in \mathbb{R}$, not vectors

solutions: (1) reduce X to a 1-dimensional score: $z = g(X) \in \mathbb{R} \Rightarrow \hat{Y} = \mathbb{1}[z \geq t]$

~~problem~~ example: body-mass-index: $z = BMI = \frac{\text{weight}}{\text{height}^2} = \frac{X_0}{X_1^2}$

[X_0: weight, X_1: height]

problem: finding good function $g(X)$ is hard => learn $g(X)$ => later

(2) reduce multi-dimensional comparison to a sequence of 1-dimensional comparisons over the elements of X ($\hat{=}$ single features) => decision tree => later

(3) nearest neighbor classifier: define the threshold implicitly via distances to "representatives" for every class => next