Classification:

- recap: - regression : Y continuous
  - classification : Y discrete
- without loss of generality: $Y = \{1, 2, \dots, C\}$ ← number of "classes" "labels"
- important special case: two-class classification, $C = 2$

$Y \in \{0, 1\}$ or $Y = \{-1, +1\}$

-1 ↑ negative, is not infected; +1 ↑ positive outcome, is infected

- objective: train a function $\hat{Y} = \hat{f}(X)$ such that $\hat{Y} = Y^*$ often (Y^* ← true resp.)

measure the quality: confusion matrix

$\hat{Y}$ \ Y^*	-1	+1
-1	true negative	false negative = missed case
+1	false positive = false alarm	true positive

$\text{TP rate} = \frac{\# TP}{\# N}$ $\quad$ $\text{TN rate} = \frac{\# TN}{N}$ $\quad$ $\text{FP rate} = \frac{\# FP}{N}$ $\quad$ $\text{FN rate} = \frac{\# FN}{N}$

$\text{total error rate} = \frac{\# FP + \# FN}{N}$ should be as small as possible

$N = TP + TN + FP + FN$

- two types of classifiers:
  - hard response: returns a single label (hopefully right)
  - soft response: returns the posterior probability $\hat{p}(Y|X)$

    $\hat{p}(Y=k|X)$ is certainty of the classifier that label k is correct

    $\hat{p}(Y=k|X) = 0.9 \Rightarrow k$ is probably the true label

    $= 0.1$ is unlikely to be the right label

    objective: return <u>well-calibrated</u> probabilities

    if $\hat{p}(Y=k|X) = 0.9 \Rightarrow$ the label k should be correct 90% of the time

    - underconfidence: classifier is right more often
    - overconfidence: classifier is right less often

Chain rule of probability theory [do not confuse with chain rule in calculus]

consider the joint probability $p(X, Y)$: the current instance has simultaneously the features X and response Y

$$\underbrace{p(X, Y)}_{\text{observe } X, Y \text{ simultaneously}} = \underbrace{p(X)}_{\text{observe } X \text{ first}} \underbrace{p(Y|X)}_{\text{observe } Y \text{, already knowing } X} = \underbrace{p(Y)}_{\text{observe } Y \text{ first}} \underbrace{p(X|Y)}_{\text{observe } X \text{ knowing } Y \text{ already}}$$

are all equivalent

$$p(X, Y, Z) = p(X)\, p(Y|X)\, p(Z|X, Y) = p(X)\, p(Z|X) \cdot p(Y|X, Z)$$
$$= p(Y)\, p(X|Y)\, p(Z|X, Y) = p(Y)\, p(Z|Y)\, p(X|Y, Z)$$
$$= p(Z)\, p(X|Z)\, p(Y|X, Z) = p(Z)\, p(Y|Z)\, p(X|Y, Z)$$

in general: $M!$ decomp. for M variables

<u>Bayes rule</u>: rewrite of the chain rule $p(X)\,p(Y|X) = p(Y)\,p(X|Y)$

$$\boxed{p(Y|X) = \frac{p(X|Y) \cdot p(Y)}{p(X)}}$$

$\left.\begin{array}{ll} p(Y|X): & \text{posterior probab.} \\ p(Y): & \text{prior probability} \end{array}\right\}$ of Y

$p(X|Y)$: likelihood of features for class Y

$p(X)$: evidence (frequency of seeing features X)

normalization: $\sum_k p(Y=k|X) = 1$ for all X

$\Rightarrow$ $p(X)$ normalizes the right-hand side

$$p(X) = \sum_k p(X|Y=k)\,p(Y=k) \Rightarrow 1 = \sum_k \frac{p(X|Y=k)\,p(Y=k)}{p(X)}$$

<u>two types of soft classifiers</u>:

- discriminative classifier: learns posterior $\hat{p}(Y|X)$ (LHS of Bayes)
- generative classifier: learns the prior $\hat{p}(Y)$ and the likelihood $\hat{p}(X|Y)$
  (term "generative": if you know the RHS of Bayes, you can create ("generate") synthetic data that are indistinguishable from real data)
- in practice:
  - discriminative cl. are usually more accurate (easier problem)
  - generative cl. are more informative (to create something, you need to understand it, harder)
- worst case performance: features are uninformative about Y

$$p(X|Y=k) = p(X|Y=k') \quad \text{for all pair } k, k' \in \{1, \dots, C\}$$

$$p(X) = \sum_k p(X|Y=k)\,p(Y=k) = p(X|Y) \underbrace{\sum_k p(Y=k)}_{=1 \text{ normalization}} = p(X|Y)$$

Bayes: $p(Y|X) = \frac{p(X) \cdot p(Y)}{p(X)} = p(Y)$ posterior equals prior

$\Rightarrow$ by observing X, we learn nothing

we can still do something ($C = 2$ for simplicity)

if $p(Y=0) < p(Y=1) \Rightarrow$ always decide for $\hat{Y} = 1$

[if $p(Y=1) \gg p(Y=0) \Rightarrow$ impressive accuracy, but false]

if $p(Y=0) = p(Y=1) = 0.5 \Rightarrow$ return $\hat{Y}$ randomly or arbitrary

$\Rightarrow$ guessing , a classifier should be at least as good as guessing

if accuracy of classifier $< 50\% \Rightarrow$ something went totally wrong

• best case: Bayes classifier : $\hat{p}(Y|X) = p^*(Y|X)$

true posterior

the best hard decision $\hat{Y} = \arg\max_k \hat{p}(Y=k|X) = \hat{f}(X)$

<u>why?</u> define the "expected error" of the classifier

$$\mathbb{E}_{X,Y^* \sim p^*(X,Y)}\left[p(\hat{f}(X) \neq Y^*)\right] = \int p(\hat{f}(X) \neq Y^*)\, \hat{p}^*(Y^*)\, \hat{p}^*(X|Y^*)\, dX\, dY^*$$

consider a particular X and two different classifier 1 and 2

such that $\hat{f}_1(X) = 0$ and $\hat{f}_2(X) = 1$

classifier 1 : $\begin{cases} \text{correct when } Y^* = 0 : \text{probability } p^*(Y^*=0|X) \\ \text{wrong when } Y^* = 1 : \quad \text{---''---} \quad p^*(Y^*=1|X) \end{cases}$ $\left.\begin{matrix} \text{wrong when } Y^* = 0 \\ \text{correct when } Y^* = 1 \end{matrix}\right\}$ classif. 2

- which of the two classifiers is preferable to minimize the error probability

  - error prob.: classifier 1 : $\hat{f}(x) = 0$ $\quad p^*(Y^* = 1 | X)$

    classifier 2 : $\hat{f}(x) = 1$ $\quad p^*(Y^* = 0 | X)$

  error is minimized when we use cl. 1 if $p^*(Y^* = 1 | X) < p^*(Y^* = 0 | X)$

  and cl. 2 otherwise

  $\Leftrightarrow$ ~~$\hat{f}(x) \neq \arg\min$~~ $\hat{f}(x) = \arg\max_{k} p^*(Y^* = k | X)$

  because then the probability of the opposite outcome is minimal

- Bayes classifier: deciding for the most probable outcome for each X also minimizes the global expected ~~error~~ error:

  $\mathbb{E}_{xy^*}[p(\hat{f}(x) \neq Y^*)] \rightarrow$ minimized when $p(\hat{f}(x) \neq Y^*)$ is minimized for every x

  if classifier has $\boxed{\hat{p}(Y|X) = p^*(Y|X)}$ and returns $\boxed{\hat{Y} = \arg\max_k \hat{p}(Y = k|X)}$ it achieves the theoretically best possible performance

  [catch: in practice, we do not know $p^*(Y|X)$, so we will at best achieve $\hat{p} \approx p^*$ $\Rightarrow$ Bayes classifier is a theoretical limit]

- for C labels, the guessing rate to beat is $\frac{1}{C}$

  this becomes harder and harder as C ~~becomes~~ increases