Machine Learning, Spring 2019: Exercise Sheet 3 (with some solutions)

Problem 1 (easy, informal). Make a list of 5 classification tasks of real-world relevance and present them in a format similar to the table at the beginning of Section 4 of the lecture notes. The purpose of this task is to make you aware of the (almost) universality of the notion of "classification" – once you start thinking of examples you'll find that many relevant real-life problems can be cast as picking the right labels for patterns.

(no solution given)

Problem 2. Consider the Digits dataset. Specify (in words or formulas) 8 binary features $f_1, \ldots, f_8$ which assign either the value 1 or the value 0 to a digit image x, such that, if you know $f_1(x), \ldots, f_8(x)$, you know which digit is shown.

Solution. For instance, put

$f_1(x) = 1$ if picture shows a strong centered vertical or slanted black line spanning the picture range, else 0. (would be a feature that gives 1 for images showing a 1, 2, or 7)

$f_2(x) = 1$ if picture contains a local patch that has a "⊣" shape, maybe a bit slanted, else 0 (indicative of digits 4, 8, 9)

$f_3(x) = 1$ if picture contains an "o" shaped substructure, else 0 (indicative of digits 0, 6, 8, 9)

etc… you get the spirit; in the end, the combination of the 8 feature values should identify a digit.

This is, by the way, the kind of hand-crafted feature design that was used by engineers in the early days of optical character recognition systems.

Problem 3. Describe a collection of binary decision questions that yield a decision tree for classifying animal classes {**Fish**, **Bird**, **Worm**, **Snake**, **Cat**, **Dog**}.

Solution. Here is a computer-science like rendering of a decision tree:

(Has_legs?	(Y Has-Wings?	(Y **Bird**)
		(N Voice=Barking?
		(Y **Dog**)
		(N **Cat**)))
	(N Has-dry-skin?	(Y **Snake**)
		(N Lives-in-Water?
		(Y **Fish**)
		(N **Worm**))))

Problem 4. (computing a decision boundary). Consider a two-class classification problem with classes c_1, c_2, where the decision is based on a single feature f: $\mathbf{P} \rightarrow \mathbb{R}$.

Let g_1, g_2 be the pdf's of the class-conditional distributions $P_{X \mid Y = ci}$ (where $i = 1, 2$). Concretely, let these pdf's be the Gaussians

$$g_1(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}, \quad g_2(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}.$$

Let the class probabilities be denoted by

$P(Y = c_i) = \gamma_i \;\; (i = 1, 2)$.

Give a formula for the decision boundary. You will find that there are three distinct cases of how the decision boundary may look like. Draw schematic plots of these three situations.

Solution. We have to solve

$\gamma_1 \, g_1(x) = \gamma_2 \, g_2(x)$

for x. Written out, this equation is

$$A_1 \, e^{-B_1 (x-\mu_1)^2} = A_2 \, e^{-B_2 (x-\mu_2)^2}$$

with $A_1 = \gamma_1 \frac{1}{\sqrt{2\pi\sigma_1^2}}, B_1 = \frac{1}{2\sigma_1^2}, A_2 = \gamma_2 \frac{1}{\sqrt{2\pi\sigma_2^2}}, B_2 = \frac{1}{2\sigma_2^2}$

Taking the log on both sides, one obtains a quadratic equation in x, which may have zero, one, or two real-valued solutions. Case "zero solutions": one of the decision functions $P(Y = c_i)\, g_i(x)$ is everywhere larger than the other. Case "one solution": the two scaled Gaussian curves intersect for only one value of the feature x. Case "two solutions": they intersect twice. See figure below for these cases. The case "one solution" is obtained if and only if $\sigma_1 = \sigma_2$ and $\mu_1 \neq \mu_2$.