Aleksandra Pizurica
E-mail: ai@lists.ugent.be

Solutions: Intro to supervised learning – Part 2

1. **Model selection and training [Old exam question]**

   Consider a classifier trained till convergence on some training data $\mathcal{D}_{train}$, and tested on a separate test set $\mathcal{D}_{test}$. You look at the test error, and find that it is very high. But when you computed the training error it was close to 0.

   (a) Which of the following is expected to help? Select all that apply.

   ☐ Increase the training data size.

   ☐ Decrease the training data size.

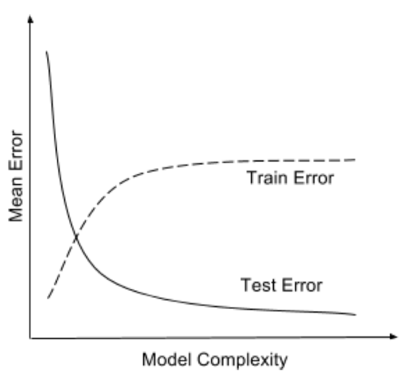
   ☐ Increase model complexity (For example, if your classifier is an SVM, use a more complex kernel. Or if it is a decision tree, increase the depth).

   ☐ Decrease model complexity.

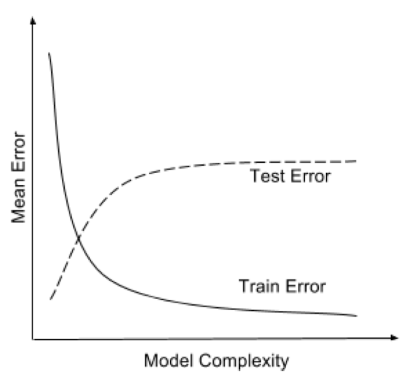
   ☐ Train on a combination of $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ and test on $\mathcal{D}_{test}$.

   (b) Explain your choices.

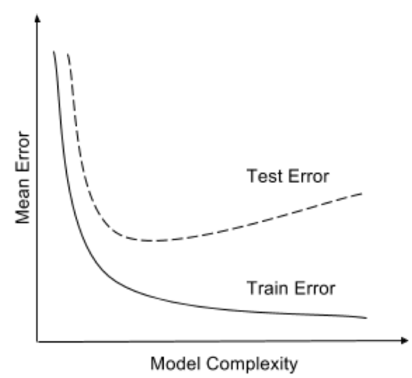
   (c) Say you plot the train and test errors as a function of the model complexity. Which of the following three plots is your plot expected to look like? Explain your choice briefly.



(i) ○



(ii) ○



(iii) ○

   (d) In machine learning, we typically need a separate validation set next to the training and test sets. What is its role?

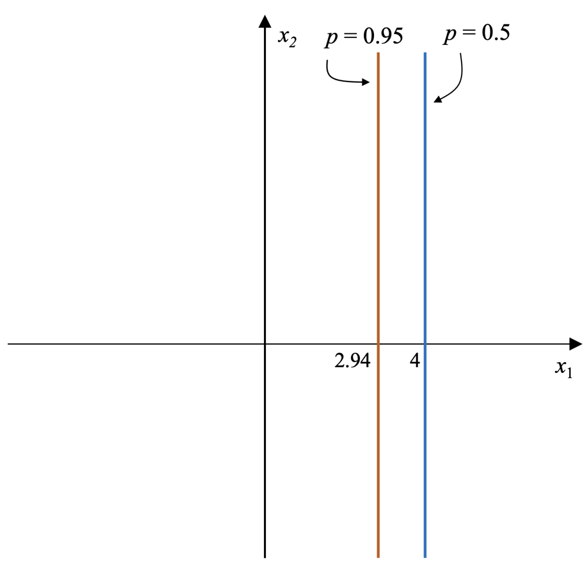
2. Consider a binary classification problem whose features are in $\mathbb{R}^2$. Suppose the predictor learned by logistic regression is $\sigma(w_0 + w_1x_1 + w_2x_2)$, where $w_0 = 4$, $w_1 = -1$ and $w_2 = 0$. Find and plot the curve along which $P(class = 1) = 0.5$ and the curve along which $P(class = 1) = 0.95$.

Solution:
The logistic function $\sigma(\mathbf{w} \cdot \mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}\cdot\mathbf{x}}}$ can be interpreted as the probability that the input sample belongs to class "1", here denoted as $class_1$. Thus, $\sigma(w_0 + w_1x_1 + w_2x_2)$ represents the probability $P(\mathbf{x} \in class_1)$.

$$\begin{aligned} \sigma(w_0 + w_1x_1 + w_2x_2) &= 0.5 \\ &\Rightarrow w_0 + w_1x_1 + w_2x_2 = 0 \\ &\Rightarrow 4 - x_1 = 0 \\ &\Rightarrow x_1 = 4 \end{aligned}$$

$$\begin{aligned} \sigma(\underbrace{w_0 + w_1x_1 + w_2x_2}_{z}) &= 0.95 \\ \sigma(z) &= \frac{1}{1+e^{-z}} = 0.95 \\ &\Rightarrow e^{-z} = -1 + \frac{1}{0.95} = 0.0526 \\ &\Rightarrow z = 2.94 \\ &\Rightarrow w_0 + w_1x_1 + w_2x_2 = 2.94 \\ &\Rightarrow 4 - x_1 = 2.94 \\ &\Rightarrow x_1 = 1.06 \end{aligned}$$

Let $p = P(\mathbf{x} \in class_1)$. The curves (which are here lines) along which p has the specified values are plotted below.



3. Consider a 3-class classification problem. You have trained a predictor whose input is $\mathbf{x} \in \mathbb{R}^2$ and whose output is $softmax(x_1 + x_2 - 1, 2x_1 + 3, x_2)$. Find and sketch the three regions in $\mathbb{R}^2$ that get classified as class 1, 2 and 3.

Solution:
The predicted class corresponds to the largest component of $softmax$, which is the same as the largest input to $softmax$.

$$\begin{aligned} z_1 &= x_1 + x_2 - 1 \\ z_2 &= 2x_1 + 3 \\ z_3 &= x_2 \end{aligned}$$

- To be classified as class 1 it needs:

$$\begin{aligned} x_1 + x_2 - 1 > 2x_1 + 3 &\quad \wedge \quad x_1 + x_2 - 1 > x_2 \\ \Rightarrow x_2 > x_1 + 4 &\quad \wedge \quad x_1 > 1 \end{aligned}$$

- To be classified as class 2 it needs:

$$\begin{aligned} 2x_1 + 3 > x_1 + x_2 - 1 &\quad \wedge \quad 2x_1 + 3 > x_2 \\ \Rightarrow x_2 < x_1 + 4 &\quad \wedge \quad x_2 < 2x_1 + 3 \end{aligned}$$

- To be classified as class 3 it needs:

$$\begin{aligned} x_2 > x_1 + x_2 - 1 &\quad \wedge \quad x_2 > 2x_1 + 3 \\ \Rightarrow x_1 < 1 &\quad \wedge \quad x_2 > 2x_1 + 3 \end{aligned}$$