# Answers Machine Learning Exercises 4

Tim van Erven

November 13, 2007

## Exercises

1. The following Boolean functions take two Boolean features $x_1$ and $x_2$ as input. The features can take on the values $-1$ and $+1$, where $-1$ represents False and $+1$ represents True. The output $y$ of the functions can also take on the values $-1$ and $+1$, with the same interpretation. For each of the functions below, either give weights for a perceptron such that the perceptron represents the function or argue that no such weights exist.

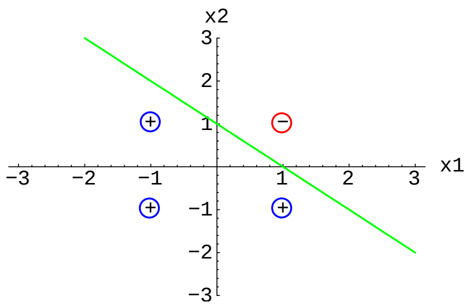
   Hint: Draw pictures like on slides 9 and 10 from mlslides8.pdf. (You do not have to submit these.)

   (a) $y = \neg\text{AND}(x_1, x_2)$

   (b) $y = \begin{cases} +1 & \text{if } x_1 = x_2 \\ -1 & \text{otherwise} \end{cases}$

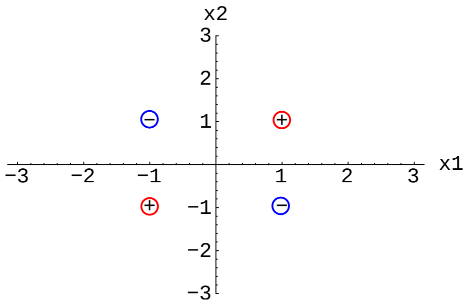
   (c) $y = \begin{cases} +1 & \text{if } x_1 = 1 \text{ and } x_2 = -1 \\ -1 & \text{otherwise} \end{cases}$

   **Answers:**

   (a) 

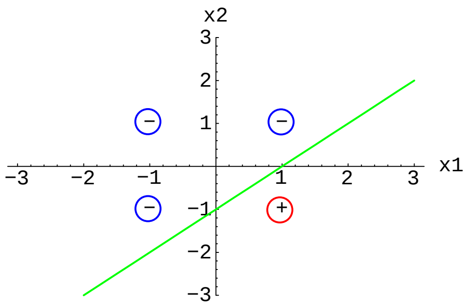

   A perceptron can represent this function using for example the weights $w_0 = 1$, $w_1 = -1$, $w_2 = -1$. Other answers are possible as well. In particular, all of these weights multiplied by the same positive constant would give the same classifications. Multiplication by a negative constant is not correct, however, because it inverts the classifications made by the perceptron.

(b)




No weights exist such that a perceptron represents this function, because the pairs of inputs and corresponding outputs are not linearly separable. (See also slide 10 of mlslides8.pdf.)

(c)



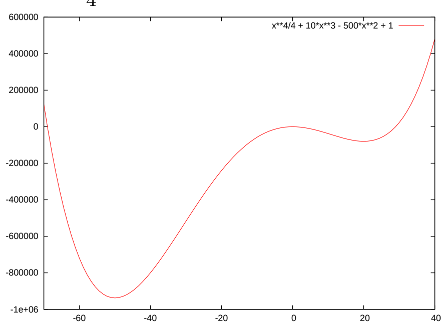

A perceptron can represent this function using for example the weights $w_0 = -1$, $w_1 = 1$, $w_2 = -1$. Again, other answers are possible as well.

**Grading:**

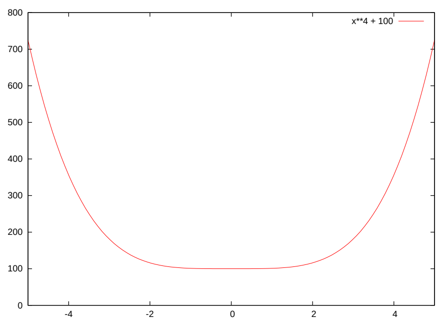
- 1 point for each correct answer.
- Giving weights that represent a correct decision boundary, but result in exactly the opposite of the desired classifications, still gives 0.5 points. For example, in (a) the answer $w_0 = -1$, $w_1 = 1$, $w_2 = 1$ would still give 0.5 points.

2. (a) For both of the following functions, argue whether gradient descent is an appropriate method to find the minimum.

$\frac{1}{4}x^4 + 10x^3 - 500x^2 + 1$




$x^4 + 100$




(b) Suppose we run gradient descent for each of the functions, regardless of whether it is appropriate. What would be $\Delta x_n$ for each of the functions when the learning rate is $\eta = 0.1$? (Work out the derivative.)

**Answers:**

(a) **N.B. A function that is not convex does not need to have local minima. It only works the other way around: If a function is convex, then it is guaranteed not to have any local minima (apart from the global one).**

$\frac{1}{4}x^4 + 10x^3 - 500x^2 + 1$**:** Gradient descent is not appropriate to find the minimum of this function, because it has a local minimum (at $x = 20$).

$x^4 + 100$**:** Gradient descent is appropriate, because $x^4 + 100$ only has one global minimum (at $x = 0$) and no other local minima. An informal argument that points this out is sufficient to get full points. For example, one might argue rather informally that $x^4$ increases faster and faster as $|x|$, the absolute value of $x$, increases, and hence it must be convex, which implies that it has no other local minima than the global minimum in the picture.
You could also have used the fact that $x^4$ is convex, which I told you during the lecture, and argued that if $x^4$ is convex, so must be $x^4 + 100$, which is just $x^4$ moved up a little.
As a third option, some of you knew that if the second derivative of a function with domain $\mathbb{R}$ is non-negative everywhere on $\mathbb{R}$, then this implies that the function is convex. This is easily verified, since

$$\frac{d^2}{dx^2}(x^4 + 100) = \frac{d}{dx} 4x^3 = 12x^2,$$

which is non-negative for any $x$.

(b) I write $x$ instead of $x_n$ to simplify the notation.

$\frac{1}{4}x^4 + 10x^3 - 500x^2 + 1$**:**

$$\begin{aligned}\Delta x &= -\eta \frac{d}{dx}(\frac{1}{4}x^4 + 10x^3 - 500x^2 + 1) \\ &= -\frac{1}{10}(x^3 + 30x^2 - 1000x) \\ &= -\frac{1}{10}x^3 - 3x^2 + 100x\end{aligned}$$

$x^4 + 100$**:**

$$\begin{aligned}\Delta x &= -\eta \frac{d}{dx}(x^4 + 100) \\ &= -\frac{4}{10}x^3\end{aligned}$$

**Grading:**

- 1 point for each of the two cases of part (a)

- 1 point for each of the two cases of part (b)

3. Suppose we have training data $D = \begin{pmatrix} y_1 \\ \mathbf{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ \mathbf{x}_n \end{pmatrix}$ and we want to use gradient descent to find weights $\mathbf{w}$ that minimize the error on $D$ for a linear unit $h_{\mathbf{w}}$. However, instead of the Sum of Squared Errors (SSE), we use a strange new error measure called the Sum of Quadratic Errors (SQE). It is defined as

$$\text{SQE}(\mathbf{w}, D) = \sum_{i=1}^{n} (y_i - h_{\mathbf{w}}(\mathbf{x}_i))^4.$$

What would be the gradient that our algorithm would use in this case? Give a derivation like in Equation 4.6 of Mitchell.

Hints: See slides 28 and 29 of mlslides8.pdf, and Equation 4.6 in Mitchell. Note that Equation 4.6 applies the chain rule, so you may have to look that up somewhere.

**Answer:** The $i$th component of the gradient is given by:

$$\begin{aligned} \frac{\partial}{\partial w_i}\text{SQE}(\mathbf{w}, D) &= \frac{\partial}{\partial w_i} \sum_{j=1}^{n} (y_j - h_{\mathbf{w}}(\mathbf{x}_j))^4 \\ &= \sum_{j=1}^{n} \frac{\partial}{\partial w_i} (y_j - h_{\mathbf{w}}(\mathbf{x}_j))^4 \end{aligned}$$

Now by the chain rule:

$$= \sum_{j=1}^{n} 4(y_j - h_{\mathbf{w}}(\mathbf{x}_j))^3 \frac{\partial}{\partial w_i}(y_j - h_{\mathbf{w}}(\mathbf{x}_j))$$

Letting $x_{jk}$ denote the $k$th component of vector $\mathbf{x}_j$, we get:

$$\begin{aligned} &= 4\sum_{j=1}^{n} (y_j - h_{\mathbf{w}}(\mathbf{x}_j))^3 \frac{\partial}{\partial w_i}(y_j - \sum_{k=0}^{d} w_k x_{jk}) \\ &= 4\sum_{j=1}^{n} (y_j - h_{\mathbf{w}}(\mathbf{x}_j))^3 (-\sum_{k=0}^{d} \frac{\partial}{\partial w_i} w_k x_{jk}) \\ &= 4\sum_{j=1}^{n} (y_j - h_{\mathbf{w}}(\mathbf{x}_j))^3 \cdot (-x_{ji}). \end{aligned}$$

Here the last equality follows because

$$\frac{\partial}{\partial w_i} w_k x_{jk} = \begin{cases} x_{jk} & \text{if } k = i, \\ 0 & \text{otherwise.} \end{cases}$$

N.B. I should have called SQE differently, because 'quadratic' means the same as 'squared' and I meant to say 'to-the-fourth'. So for example "Sum of Strange Errors" would have been better.

**Grading:**

- 2 points for a correct answer.

## Grading Policy

- Grades are between 1 and 10.
- You always start with 1 point.
- Partial points may be awarded for partially correct answers.