Machine Learning and Pattern Recognition
Tutorial Sheet Number 4 Answers

School of Informatics, University of Edinburgh, Instructor: Chris Williams

1. Let $E(\boldsymbol{w})$ be a differentiable function. Consider the gradient descent procedure

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t - \eta \nabla_{\boldsymbol{w}} E$$

Let's say we initialize $\boldsymbol{w}^0 = \mathbf{0}$. True/false:

(a) Let $\boldsymbol{w}^1$ be the result of taking one gradient step. Then the error always improves, i.e., $E(\boldsymbol{w}^1) \leq E(\boldsymbol{w}^0)$.

(b) There exists some choice of the step size η such that $E(\boldsymbol{w}^1) < E(\boldsymbol{w}^0)$.

Solution:

(a) False. If η is too large we may increase the value of E even though we are going in the negative gradient direction.

(b) False. If were at a local optimum, then the gradient is zero and gradient descent will not change the location. However, this is a bit of a trick question in that if $\nabla_{\boldsymbol{w}} E \neq 0$, then the statement is indeed true: If we take a small enough step, we're guaranteed to decrease the function. Using a Taylor expansion, we have that

$$E(\boldsymbol{w}^t + \Delta\boldsymbol{w}) = E(\boldsymbol{w}^t) + (\Delta\boldsymbol{w})^T \nabla_{\boldsymbol{w}} E + \ldots \tag{1}$$

Setting $\Delta\boldsymbol{w} = -\eta \nabla_{\boldsymbol{w}} E$ we have that

$$E(\boldsymbol{w}^t + \Delta\boldsymbol{w}) = E(\boldsymbol{w}^t) - \eta (\nabla_{\boldsymbol{w}} E)^T \nabla_{\boldsymbol{w}} E + O(\eta^2) \tag{2}$$

so the change is negative for small enough η.

■

2. In the gradient descent procedure, a common programming mistake is to forget the minus sign, i.e., you unintentionally write a procedure that does

$$\boldsymbol{w}^{t+1} \leftarrow \boldsymbol{w}^t + \eta \nabla_{\boldsymbol{w}} E$$

If you make this mistake, what happens?

Solution: The algorithm *maximizes* E rather than minimizing it. Ordinarily, if you really did mean to minimize E, then it will have no maximum. E.g., in linear regression, it is possible to make the squared error as large as you like by using a particularly poor set of weights. So typically you will see $E(\boldsymbol{w}^t)$ go to infinity as $t \to \infty$. ■

3. Consider the following classification problem. There are two real-valued features $x_1, x_2 \in \mathbb{R}$ and a binary class label. The class label is determined by

$$y = \begin{cases} 1 & \text{if } x_2 \geq |x_1| \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

(a) Can this be perfectly represented by a feedforward neural network without a hidden layer? Why or why not?

(b) Let's consider a simpler problem for a moment. Consider the classification problem.

$$y = \begin{cases} 1 & \text{if } x_2 \geq x_1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Design a single neuron that solves this problem. Pick the weights by hand. For an activation function, use the hard threshold function

$$h(a) = \begin{cases} 1 & \text{if } a \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

(c) Now go back to the classification problem at the beginning of this question. Design a two layer feedforward network (i.e., one hidden layer) that represents this function. Use the hard threshold activation function as in the previous question. *Hints:* Use two units in the hidden layer. The unit from the last question will be one of the units, and you will need to design one more. Your output unit will essentially perform a binary AND operation on the hidden units.

Solution:

(a) No. A neural network without a hidden layer simply computes a linear function of the inputs. The decision boundary required for this problem is not linear.

(b) Call the neuron z_1, and compute its output by

$$z_1 = h(v_{11}x_1 + v_{12}x_2 + v_{10})$$

Set

$$v_{11} = -1, v_{12} = 1, v_{10} = 0.$$

Now $z_1 > 0 \iff x_2 - x_1 \geq 0$.

(c) Define another hidden unit z_2 as

$$z_2 = h(v_{21}x_1 + v_{22}x_2 + v_{20})$$

set

$$v_{11} = 1, v_{12} = 1, v_{10} = 0.$$

Now $z_2 > 0 \iff x_1 + x_2 \geq 0$. So $x_2 \geq -x_1$.

The intersection of the area for which $z_1 \geq 0$ and the area for which $z_2 \geq 0$ is the area where $y = 1$. We can set up a logical AND with the following output unit:

$$f = h(w_1 z_1 + w_2 z_2 + w_0),$$

where we choose

$$w_1 = 1, w_2 = 1, w_0 = -1.1$$

You can show that now $y = 1 \iff x_2 \geq |x_1|$. Try this on a few example points to verify that it works.

■

4. The following problem was introduced in the lectures. Work through the rest of the problem to obtain the solution.

You are an auditor of a firm. You receive details about the sales that a particular salesman is making. He attempts to make 4 sales a day to independent companies. You receive a list of the number of sales

by this agent made on a number of days. Explain why you would expect the total number of sales to be binomially distributed.

If the agent was making the sales numbers up as part of a fraud, you might expect the agent (as he is a bit dim) to choose the number of sales at random from a uniform distribution. You are aware of the fraud possibility, and you understand there is something like a 1/5 chance this salesman is involved. Given daily sales counts of 1 2 2 4 1 4 3 2 4 1 3 3 2 4 3 3 2 3 3, do you think the salesman is lying?

From the lectures we had:

- $\mathcal{M} = 1$ - From $P_1(x|p)$ a binomial distribution Binomial$(4, p)$. Prior on p is uniform.
- $\mathcal{M} = 2$ - From $P_2(x)$ a uniform distribution Uniform(0,...,4).
- $P(\mathcal{M} = 1) = 0.8$.

$$P(\mathcal{D}|\mathcal{M} = 1) = \int dp\ P_1(\mathcal{D}|p)P(p)\ ,\ P(\mathcal{D}|\mathcal{M} = 2) = P_2(\mathcal{D})$$

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D}|\mathcal{M} = 1)P(\mathcal{M} = 1) + P(\mathcal{D}|\mathcal{M} = 2)P(\mathcal{M} = 2)}$$

To get the solution, you should look at the form of the density of a beta distribution and note that it must integrate to 1. You also will need to be able to compute with Γ functions of big numbers. In matlab it is worth doing the computations in log space, and using the `gammaln` function. Then you can exponentiate it when you have done all the sums.

Solution: Look up the Binomial Distribution. For a single observation of a count $X = r$ under a Binomial distribution with parameter p and n repetitions we have

$$p(X = r|p) = \binom{n}{r} p^r (1-p)^{(n-r)}.$$

We can thus write the likelihood out as

$$P_1(\mathcal{D}|p) = K p^{50} (1-p)^{26}$$

where $K = 4^3 6^5 4^7$. This follows from filling in the numbers into the product of binomials, assuming the data points are IID, and by counting the numbers of each case (3 ones, 5 twos, 7 threes and 4 fours. $3 \times 1 + 5 \times 2 + 7 \times 3 + 4 \times 4 = 50$. 19 data points. Maximum possible total count $4 \times 19 = 76 = 50 + 26$). Note also

$$\binom{4}{1} = 4, \qquad \binom{4}{2} = 6, \qquad \binom{4}{3} = 4, \qquad \binom{4}{4} = 1.$$

Thus likelihood has a form of a scaled beta distribution, so, with prior $P(p) = 1$ we can work out the integral:

$$\int_0^1 dp\ p^{50}(1-p)^{26} = \frac{\Gamma(51)\Gamma(27)}{\Gamma(78)}$$

from looking at the normalisation constant for the Beta distribution.

Hence the $P_1(\mathcal{D})$ is

$$4^3 6^5 4^7 \frac{\Gamma(51)\Gamma(27)}{\Gamma(78)}$$

To compute this we need to use log space and the `gammaln` function. The resulting marginal likelihood is 6.89×10^{-13}.

For the uniform distribution $P_2(\mathcal{D})$ there are options of $0, 1, 2, 3, 4$ sales. Hence the uniform probability is 0.2 for each possibility for each data point. 19 data points results in a likelihood of $0.2^{19} = 5.25 \times 10^{-14}$.

Now combine this with the priors for the two models (0.8 and 0.2), and the posterior probablity is about 0.98 in favour of the salesman being an honest chappie.

■