Exercise 10: Softmax and Bayes' Equation

Lecture Information Processing and Communication

Jörn Anemüller, June 2022

Submit solutions until Tuesday 2022-06-28, 23:59h, by uploading to your group's exercise folder on cs.uol.de. You may submit your solutions in groups of at most two students.

Note that you will receive the next programming exercise by Friday, June 24rd, for which you will have time to finish until the week after.

1. Softmax-Function

(a) First, consider logistic regression. Show that the logistic function $g(z) = 1/(1+\exp(-z))$ computed for a value of z and a value of $-z$ adds up to a value of one:

$$1 \equiv g(z) + g(-z)$$

Therefore, in a logistic regression classification task with a weight vector $\mathbf{w}$, the probability estimates for the two classes may be expressed as

$$P(class_1|\mathbf{w}, \mathbf{x}) = g(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$$
$$P(class_0|\mathbf{w}, \mathbf{x}) = g(-\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})}$$

and represent properly normalized probabilities.

(b) Now consider a simple linear neural network with two output units that are used (via one-hot encoding) to discriminate between two classes. Assume that the outputs are scaled with the softmax function σ according to

$$[\sigma(\mathbf{z})]_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

and the inputs are processed with a single linear mapping

$$\mathbf{z} = \mathbf{W}\mathbf{x}.$$

Consider the special case of the weight matrix being composed of two identical but sign-reversed row vectors, i.e.,

$$\mathbf{W} = \begin{pmatrix} \mathbf{w}^T \\ -\mathbf{w}^T \end{pmatrix}.$$

Show that the probability estimates of this network are identical to the logistic regression probability estimate from (a), provided that a trivial rescaling of the weight vectors is applied. Which rescaling transforms the weight vector from part (a) above into weight vectors in part (b) that result in the same probability estimates?

2. Derivation of Bayes' equation

(a) Explain the terms joint probability (Verbundwahrscheinlichkeit) $P(x, y)$, marginal probability (Marginalwahrscheinlichkeit) $P(x)$ and $P(y)$, respectively, and conditional probability (bedingte Wahrscheinlichkeit) $P(x|y)$ and $P(y|x)$, respectively.

(b) Rewrite the joint probability $P(x, y)$ in a convenient way and derive Bayes' equation from it.

(c) Indicate what the terms "prior probability", "likelihood", and "posterior probability" refer to in the context of Bayes' equation.

(c) The joint probability of two events R ("rain") und S ("sun") is given by the values according to the table below:

P(R,S)	R=0	R=1
S=0	0.4	0.2
S=1	0.3	0.1

Check that the probabilities are properly normalized, i.e., that the sum over the probablities for all possible event-pairs evaluates to $1.$;)
Compute the values of the marginal probabilities for R and S, as well as the conditional probabilities $p(R|S)$ und $p(S|R)$.
Check the validity of Bayes' equation for the case of $R = 1$ and $S = 1$.

3. Priors and posteriors

Joe gets tested for a nasty disease. Let the true state of Joe's health be denoted by variable a, and the test result by b.

$$a = 1 \quad \text{Joe has the disease}$$
$$a = 0 \quad \text{Joe does not have the disease}$$

The test result is either "positive" ($b = 1$) or "negative" ($b = 0$). The test is 95% reliable: In 95% of cases of persons who do have the disease, the test will give a positive result; and in 95% of cases of persons who do *not* have the disease, it will give a negative result. It is known that in Joe's age group, 1% of all persons have the disease.
Joe's test result was positive. How large is the probability that he does actually have the disease?