# Machine Learning II – Assignment 1

Stefan Roth, Tobias Plötz, Uwe Schmidt

October 22, 2014

This homework is due on November 3, 2014 at 20:00.
**Please read the instructions carefully!**

## General remarks for all assignments

Your grade will of course depend on the correctness of your answer, but also on a clear presentation of your results and good writing style. It is your responsibility to find a way to *explain clearly how* you solved the problems. Note that you will get grades for the solution, not for the result. If you get stuck, try to explain why and describe the problems you encountered – you can get partial credit even if you did not complete the task. So please hand in enough information for us to understand what you did, what you tried, and how it worked!

We encourage interaction about class-related topics both within and outside of class. However, you should not share solutions with your classmates, and *everything you hand in must be your own work*. You are also not allowed to use material from the web. You are required to **acknowledge any source of information that you used to solve the homework** (i.e. books other than the course books, papers, web sites, etc.). Acknowledgements will *not* affect your grade. Thus, there is no reason why you would not acknowledge sources properly. Not acknowledging a source that you have used, on the other hand, is a clear violation of academic ethics. Note that the university as well as the department is very serious about plagiarism. For more details please see `http://www.informatik.tu-darmstadt.de/index.php?id=202` and `http://plagiarism.org`.

### Groups

Please form groups of two people to work on the exercises and do not change groups during the semester. If you do not find a group member get in touch with us *early* and we will assign you to another person.

### Programming exercises

For the programming exercises you will be asked to hand in Matlab code. If you used any other tool to write your code, say Octave, it is *your responsibility* to make sure that the code also works in Matlab, which is what we will use for grading. We have Matlab access available for everyone who needs it. In order for us to be able to grade the programming assignments properly, you need to comment your code in sufficient detail so that it will be easily clear to us what each part of your code does. Sufficient detail does not mean that you should comment every line of code (that defeats the purpose), nor does it mean that you should comment 20 lines of code using only a single sentence. Of course, all this is good coding practice anyway, so you would want to do this no matter what we expect from you.

Your Matlab code should display your results so that we can judge if your code works from the results alone. Of course, we will still look at the code. If your code displays results in multiple stages, please insert appropriate `pause` commands between the stages so that we can step through the code. Please be sure to name each file according to the naming scheme included with each problem. This also makes it easier for us to grade your submission. And finally, please be sure to include your name and email in the code.

### Files you need

All the data you will need for the problems will be made available on Moodle `https://moodle.tu-darmstadt.de/course/view.php?id=4197` and on the course web page `http://www.gris.tu-darmstadt.de/teaching/courses/ws1415/ml2/`.

### What to hand in

As mentioned, you need to show your solution and how you got there. Your handin should contain a PDF file (plain text is ok, too) with any textual answers that may be required; also put your name on the first page. You do not have to include images of your results. Your code should show these instead.

For the programming parts, please hand in all documented `.m` scripts and functions that your solution requires. Make sure your code actually works (also in an empty workspace) and that all your results are displayed properly.

### Handing in

Please upload your writeup and your code to the corresponding Moodle area: `https://moodle.tu-darmstadt.de/course/view.php?id=4197`. Only one group member has to hand in your work. If *and only if* you experience problems with uploading your solution, you may also email it to `ml2staff@gris.tu-darmstadt.de`

You are supposed to send all your solution files as a single `.zip` or `.tar.gz` file. **Please note that we cannot accept file formats other than the ones specified!** These are widespread standards that are available on any platform.

### Late Handins

We will accept late handins, but we will take 20% off for every day that you are late. Note that even 15 minutes late will be counted as being one day late! After the exercise has been discussed in class, you can no longer hand in.

If you, for some serious (say medical) reason, cannot make the deadline, you need to contact us *before* the deadline. We might waive the late penalty in such a case.

## 1 Bayesian Decision Theory (10 points)

In a typical classification or regression problem we want to predict the "best" outcome $\hat{y}$ given a value for an observed data point $x$. According to Bayesian decision theory we should pick $\hat{y}$ such that the expected loss with respect to a loss function $\Delta$ and the posterior probability $p(y|x)$ is minimized, i. e. for continuous $y \in \mathbb{R}$

$$\hat{y} = \operatorname*{arg\,min}_{y' \in \mathbb{R}} \int_{\mathbb{R}} \Delta(y', y) p(y|x) dy \tag{1}$$

and for $y$ from some discrete set $\mathcal{Y}$

$$\hat{y} = \operatorname*{arg\,min}_{y' \in \mathcal{Y}} \sum_{y \in \mathcal{Y}} \Delta(y', y) p(y|x) \tag{2}$$

Please derive closed form expressions for $\hat{y}$ given the following loss functions:

1. The squared loss for continuous $y \in \mathbb{R}$: $\Delta(y', y) = (y' - y)^2$ .
2. The 0/1 loss for discrete $y \in \mathcal{Y}$: $\Delta(y', y) = 1 - \delta(y', y)$, with $\delta$ denoting the Dirac delta, i. e. $\delta(y', y) = 1$ iff $y' = y$ and 0 else.

## 2 Decision Regions (5 points)

Consider two nonnegative numbers $a$ and $b$ and show that, if $a \leq b$, then $a \leq (ab)^{1/2}$. The probability of misclassification is given by

$$p(\text{mistake}) = \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}. \tag{3}$$

Assume that the decision regions of a two-class classification problem are chosen to minimize the probability of misclassification, which means that each $\mathbf{x}$ is assigned to whichever class has the smaller value of the integrand in (3). Show that as a consequence, the probability will satisfy

$$p(\text{mistake}) \leq \int \{p(\mathbf{x}, \mathcal{C}_1) p(\mathbf{x}, \mathcal{C}_2)\}^{1/2} d\mathbf{x}. \tag{4}$$

## 3 Probabilities (5 points)

1. Show that

$$p(x, y|z) = p(x|z) p(y|x, z), \tag{5}$$

and also

$$p(x|y, z) = \frac{p(y|x, z) p(x|z)}{p(y|z)}. \tag{6}$$

2. The following famous problem is known as the three door problem: In a game show a candidate is offered to choose from three doors. Behind one door there is a car while behind the other two doors there is a Zonk. The candidate would like to win the car! Now the candidate can choose one of the three doors. After he chose and told the game-show host, one from the other two doors **not** containing the car is opened. The game-show host then offers the candidate to reconsider his choice and allows him to select from the two remaining doors. The question is whether the candidate should stick with his initial choice, or choose the other remaining door. Of course it is better to switch doors. You are the candidate, and you are a Bayesian. Explain why you should switch using the notion of random variables and the terms prior, posterior and conditional probability. (Hint: Use three RVs, one for your choice, one for the host's choice and one for the car)