School of Electronics and Computer Science
University of Southampton

COMP6229 (2017/18): Machine Learning Lab 1 Not for assessment

Issue	2 Oct. 2017
Deadline	10 Oct. 2017

This exercise supplements material taught in the lectures. It is a mandatory part of the module, but is not for assessment; spend about 10 hours on it in timetbled lab sessions and afterwards. If you are unfamiliar with MATLAB, you may spend more time to become skilled at it. We assume that at this level (Part III / MSc) if you are competent in one high level language, picking up the basics of a scripting language should be straightforward.

$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{C}),\ \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \implies \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{m},\ \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^T)$

1. Familiarize yourself with MATLAB. Work through the examples in the document `http://users.ecs.soton.ac.uk/mn/MatlabIntroduction.pdf`, which are notes accompanying a textbook on MATLAB. Use of the `help` and `lookfor` commands help you learn a broad range of the features of the language. The documents `http://users.ecs.soton.ac.uk/mn/MatlabProgramming.pdf` and `http://users.ecs.soton.ac.uk/mn/MatlabStyle.pdf` are also worth going through, but not essential to get started.

2. Generate 1000 uniform random numbers and plot a histogram. Here are the useful commands in MATLAB to do this.

```
> x = rand(1000,1);
> hist(x,40);
> help hist
> [nn, xx] = hist(x);
> bar(nn);
```

   Repeat the above with 1000 random numbers drawn from a Gaussian distribution of mean 0 and standard deviation 1 using `x = randn(1000,1);`. You now see how data drawn from two different probability densities are distributed. Change the number of bins into which data is split (*i.e.* the 40 in `hist(x,40)`) and note the differences.
   Now try the following

```
> N = 1000;
> x1 = zeros(N,1);
> for n=1:N
>   x1(n,1) = sum(rand(12,1))-sum(rand(12,1));
> end
> hist(x1,40);
```

   What do you observe? Is there a theorem that explains your observation?
   Note: Do not type the MATLAB statements one at a time into command line; type them into a text file `labone.m` and invoke the script by `> labone` against the prompt. Look up `> help path`.

3. Consider the covariance matrix $\boldsymbol{C} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.
   Factorize this into $\boldsymbol{A}^t\boldsymbol{A} = \boldsymbol{C}$ using `> A = chol(C)`.
   Confirm the factorization is correct by multiplying. Generate 1000 bivariate Gaussian random numbers by `X=randn(1000,2);`
   Transform each of the two dimensional vectors (rows of `X`) by `> Y=X*A`.

Now draw a scatter plot of `X` and `Y`.
`> plot(X(:,1),X(:,2),'c.', Y(:,1),Y(:,2),'mx');`
What do you observe?

Construct a vector $\boldsymbol{u} = [\sin\theta \ \cos\theta]$, parameterized by the variable θ and compute the variance of projections of the data in $\boldsymbol{Y}$ along this direction:

```
> theta = 0.25;
> u = [sin(theta); cos(theta)]
> yp = Y*u;
> var_empirical = var(yp)
> var_theoretical = u'*C*u;
```

In the above, the variance of projections has been calculated in two ways (theoretical, *i.e.* from a formula and empirical, *i.e.* by simulating data). Explore how the difference between the two changes with the number of data points used (at 10, 100, 1000 and 10000).
Plot how this projected variance changes as a function of θ:

```
> N = 50;
> plotArray = zeros(N,1);
> thRange = linspace(0,2*pi,N);
> for n=1:N
> ...
> ...
> end
> plot(plotArray)
```

Explain what you observe by calculating the eigenvectors of the covariance matrix.

Derive an expression for the projected variance analytically for this two-dimensional case (*i.e.* the variance of projected data as a function of θ) and confirm that the plot you have drawn is correct.

How does what you have done above differ for $\boldsymbol{C} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

Export the figures for inclusion in a report `> print -depsc f1.eps`.

4. Describe the work you have done as a short report. Upload a *pdf* file **no longer than two pages** (two pages absolute maximum, no cover pages / appendices) using the ECS handin system: `http://handin.ecs.soton.ac.uk`. If possible, please use LaTeX to typeset your report. Please make sure your name and email are included.

Important Note:

- You have to work independently on this and future assignments. This module does **not** encourage group working.
- During timetabled laboratory sessions, the working language is English, both for communicating with staff and among students.