# Assignment-1 (CS-419)

Due Date: 2-Mar-2014

**Note:** Please do not copy answers from your friends. This assignment carries NO explicit marks. However it is mandatory to answer ALL questions to pass the course. You are free to use any programming language or software or scripting language. I recommend using SCILAB or MATLAB. Every question has a clearly mentioned deliverable(s). You MUST create a report consisting of ONLY these deliverables. Do not write anything other than the deliverables in the report. Upload the report in your favourite cloud storage (like dropbox or skydrive) and email the link ALONE to your TA: paramita2000@gmail.com . She may additionally ask some of you to demo the scripts/programs you wrote for creating the deliverables. Failure to demo will essentially prove that you copied and will lead to a fail grade in the course and registering a complaint to the disciplinary committee.

1. Consider the linear model and the ideal goal of minimizing risk with squared-loss:

$$\mathbf{w}^*_{TRM} \equiv \text{argmin}_{\mathbf{w}\in\mathbb{R}^n} \mathbb{E}\left[\left(\mathbf{w}^\top X - Y\right)^2\right].$$

   For this case we discussed an algorithm, known as Empirical Risk Minimization (ERM):

$$\mathbf{w}^m_{ERM} \equiv \text{argmin}_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{w}^\top \mathbf{x}_i - y_i\right)^2,$$

   which is Probably Aproximately Correct (PAC) or in other words statistically consistent. Your goal in this exercise is to observe PAC-ness empirically. Here is exactly what you have to do:

   - Assume that the unknown distribution $U_{XY}$ is a (multivariate) Gaussian with mean $\mu \in \mathbb{R}^{n+1}$ and covariance matrix $\Sigma \succ 0$ of size $n+1$. Assume the first $n$ dimensions generate $X$ and the last one generates $Y$. Derive analytical expressions for the minimizer of the true risk ($\mathbf{w}^*_{TRM}$) as well as the empirical risk ($\mathbf{w}^m_{ERM}$) in terms of $\mu, \Sigma$ and their sample estimates respectively.

- Choose[1] a particular $\mu, \Sigma$ and from $U_{XY}$ generate[2] $p$ training sets each with $m$ sample pairs $(\mathbf{x}_i, y_i)$. Lets call them $\mathcal{D}_1^m, \ldots \mathcal{D}_p^m$. Now repeat this exercise of creating $p$ training sets with various $m$, say $m = 10, 50, 100, 500, 750, 1000, 2500, 5000, 7500, 10000$, while keeping the chosen mean and covariance the same. Make sure $\mathcal{D}_i^{m_1} \subset \mathcal{D}_i^{m_2}$, $\forall i$, whenever $m_1 < m_2$.
- Using the analytical expressions derived, actually compute the minimizers $\mathbf{w}^*_{TRM}$ and for all training sets $\mathbf{w}^*_{ERM}$ (numerically).
- Let the fraction of training sets where $\|\mathbf{w}^*_{TRM} - \mathbf{w}^*_{ERM}\| > 10^{-3}$ be $f$.
- For $n = 1, 5, 10, 50, 100$ provide plots[3] of $f$ vs. $m$. Do your plots provide an evidence for statistical consistency?

The deliverables for this question are: i) the values of $\mu, \Sigma$ you pick and, ii) the five plots.

2. Consider a die with six faces and true probabilities $P[X = 1] = P[X = 3] = P[X = 5] = \frac{1}{12}$ and $P[X = 2] = P[X = 4] = P[X = 5] = \frac{1}{4}$. Now using this true multinoulli distribution generate[4] 6000 training sets: $\mathcal{D}_m = \{x_1, \ldots, x_m\}$ with $m = 1, 2, \ldots, 6000$, where $x_i$ is the face shown in the $i^{th}$ roll of this die. Note that $\mathcal{D}_1 \subset \mathcal{D}_2 \subset \ldots \subset \mathcal{D}_{6000}$. Consider three models for learning the true multinoulli distribution using the datasets: i) Multinoulli ii) Dirichilet-Multinoulli with hyperparameters[5] that imply that most-likely the probabilities of odd faces are thrice that of the even faces iii) Dirichilet-Multinoulli with hyperparameters that imply that most-likely the probabilities of even faces are thrice that of the odd faces. For the first model employ MLE and the other two consider two alternatives: MAP and BAM. Hence there are totally five algorithms. Let $\theta^*, \theta^m_{A1}, \ldots, \theta^m_{A5}$ be the vectors of probabilities of the six faces with the die and those obtained with the five algorithms using $\mathcal{D}_m$ respectively. For each algorithm $Ai, i = 1, 2, 3, 4, 5$, plot $\|\theta^* - \theta^m_{Ai}\|$ vs. $m$. Is your observation from these plots intuitive?

   The deliverables for this problem are: i) the hyperparameters you pick and, ii) the five plots.

[1] If two friends pick the same $\mu, \Sigma$ this will imply they copied from each other ;)

[2] Using computer simulation. If two friends use similar similar scripts, then they have copied ;).

[3] You may use more number of $m$ values to get a better resolution plot.

[4] Using computer simulation.

[5] Note that there are multiple choices of hyperparameters that imply this. You can pick any one of them. If two friends pick the same hyperparameters then that will imply that they copied answers ;)