8 Machine Learning and Bayesian Inference (sbh11)

(*a*) State the *central limit theorem* for a sequence $X_1, X_2, \ldots, X_n$ of independently and identically distributed (iid) random variables having mean $\mathbb{E}(X) = \mu$ and variance $\text{var}(X) = \sigma^2$. [2 marks]

(*b*) Explain how the central limit theorem can be used to provide a *two-sided confidence interval* for the estimate of the mean of a random variable. [4 marks]

(*c*) When $X \in \{0, 1\}$, explain how an estimate of $\mathbb{E}(X)$ can be used to obtain an estimate of $\text{var}(X)$. [2 marks]

(*d*) We know that if X is normal distributed with mean 0 and variance 1 then $\Pr(-1.96 \leq X \leq 1.96) > 0.95$. You have trained a classifier using algorithm A on a data set and tested it using 1000 test examples. You obtain 57 errors. Find a two-sided 95% confidence interval for the accuracy. [3 marks]

(*e*) You have a trained second classifier using algorithm B and the same data set as in Part (*d*), which, when tested using the same 1000 test examples, gives 55 errors. Denoting by a the actual accuracy of the first classifier, and by a' the actual accuracy of the second classifier, find a two-sided 95% confidence interval for the difference $(a - a')$ in the accuracies of the two classifiers. State any new assumptions that you make. [4 marks]

(*f*) Your boss has a vested interest in arguing that algorithm B is better than algorithm A for the problem of interest, and argues that the measured accuracies confirm this view. Discuss the pros and cons of this conclusion, and suggest how you might conduct further experiments to help your boss. [5 marks]