CO902
Problem Set 3

1. *Bounding a classification error rate.* For an arbitrary classification problem, let X_i be 1 if the ith case is correctly predicted, 0 otherwise, $i = 1, \ldots, n$. Note the change in notation from previous usage: X_i is *not* the data; it is the end result of your classifier applied to case i (in the supervised setting where you know the true class). Regard the X_i as independent Bernoulli random variables; then $\mathsf{E}(X_i) = \theta$ is the true–but unknown–error rate of your classifier. Consider the MLE estimator of θ.

   (a) Use Chebyshev's inequality (see previous problem set) to find a bound on the probability that the MLE estimator is farther than a units away from the true error rate.

   (b) Thinking of the Spam example where the accuracy was around 85%, compute this bound for $\theta = 0.85$, $a = 0.085$ (10% of θ) and for n of (at least) 10, 100, & 1000. Argue for a particular n as being "enough" in that example based on this result.

   (c) There is a fudge here. What assumption is questionable? How might this disturb our results?

2. *Bernoulli MAP properties.* Reconsider (the much considered) Bernoulli example: $X_i \sim \text{Ber}(\theta)$, iid, $i = 1, \ldots, n$.

   (a) Show that the MLE of θ is unbiased and consistent.

   (b) Consider the MAP estimator when a Beta prior is used, $\theta \sim \text{Beta}(\alpha, \beta)$. Is the MAP estimator unbiased? Is it consistent? (Work with arbitrary α & β, i.e. don't use specific values as we've done in class and lab.)

3. *Bayes for Gaussian random variables.* Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ independently, and for simplicity assume that σ^2 is known. Use a Normal prior for the mean, $\theta \sim N(a, b^2)$. Show that the posterior is

$$\theta | X_1, \ldots, X_n \quad \sim \quad N(\bar{\theta}, \tau^2)$$

   where

$$\begin{aligned} \bar{\theta} &\sim w\bar{X} + (1-w)a \\ w &= \frac{\frac{1}{\mathsf{se}^2}}{\frac{1}{\mathsf{se}^2} + \frac{1}{b^2}} \\ \frac{1}{\tau^2} &= \frac{1}{\mathsf{se}^2} + \frac{1}{b^2} \\ \mathsf{se} &= \frac{\sigma}{\sqrt{n}} \end{aligned}$$

Inverse variance is also known as precision; what role does the prior precision and data precision play in w? What role doe they play in the posterior precision? What happens as n grows?

4. *Iterated Expectation & Variance.* Prove these two fundamental and incredibly useful identities. For continuous random variables X and Y with joint distribution $p(x, y)$ and marginal distributions $p(x)$ and $p(y)$,

$$\begin{aligned} \mathsf{E}(X) &= \mathsf{E}_Y\left(\mathsf{E}_X(X|Y)\right) \\ \mathsf{Var}(X) &= \mathsf{E}_Y\left(\mathsf{Var}_X(X|Y)\right) + \mathsf{Var}_Y\left(\mathsf{E}_X(X|Y)\right) \end{aligned}$$

The subscripts remind you how to compute each expectation or variance. In the expression $X|Y$, the random variable Y is "conditioned upon" and thus there is no randomness in Y, and thus any expectation must be w.r.t. X alone (E_X). For $\mathsf{E}_X(X|Y)$, note that an expectation with respect to X integrates out all dependence on X (i.e. the integral $\int \cdot\, p(x)dx$ cannot depend on x), and so $\mathsf{E}_X(X|Y)$ can only depend on Y (and thus E_Y). The same goes with the second expression with the variances.

Note that these expressions are actually unambiguous with out the subscripts—there is only one way you can evaluate them—but the subscripts help 'lead the way'.