Announcements:

- This is the last hw, and the last assignment of any kind, for the course!
- Please don’t forget to do course evaluations at http://www.cmu.edu/fce.
- There will be NO CLASS Fri May 5, when I will be in DC (I will actually be leaving Thur May 4).
- The material for the last part of the course comes from Chapter 4 and Sections 12.1–12.3 (discriminant analysis and support vector machines); Sections 8.5 and 14.3 (cluster analysis); and Sections 14.5–14.7 (however much of this I get to!).

Problems:

1. Please do Exercise 4.6, p. 113. This proves that the perceptron algorithm converges in a finite number of steps when the classes are linearly separable. It might also be worth re-viewing lec31.r from the week13 area of the website.

2. Kernelized linear regression. The “kernel trick” in svm’s was to replace inner products $x_i^T x_j$ in the original feature space with inner products $h(x_i)^T h(x_j) = K(x_i, x_j)$ in a new feature space. Then we only have to compute with the kernel function $K(\cdot, \cdot)$ to obtain an svm. Section 12.3.6, pp. 387–389 of HTF, shows how to do the same thing for penalized linear regression.

Please do Exercise 12.3, p 406 of HTF, which considers a modification of kernelized regression, in which the intercept is left unpenalized.

Note (not required for this exercise): Other prediction/classification functions can be kernelized also. For example, Exercise 12.10, p. 408 of HTF, shows how kernelize LDA.

3. Consider the $K$-component normal mixture model for observations $y_1, \ldots, y_n \in \mathbb{R}^d$

$$p(y_j \mid \phi) = \sum_{k=1}^{K} \alpha_k n_d(y_j \mid \mu_k, \sigma^2 I)$$

where $\alpha_k \geq 0$, $\sum_k \alpha_k = 1$, and $I$ is the $d \times d$ identity matrix.

(a) Write down the observed-data log-likelihood.

(b) Derive an E-M algorithm for computing the MLE of $\phi = (\alpha, \sigma^2)$. How does the algorithm change if $\sigma^2$ is fixed and known (and so doesn’t need to be estimated)?

(c) Show that if $\sigma$ has a known value, and we take $\sigma \rightarrow 0$, then the E-M algorithm in a sense coincides with $K$-means clustering.

4. Refer again to parts (a) and (b), and the hierarchical model stated in part (c), of problem #2, HW 04.

(a) Develop the E and M steps of an E-M algorithm for computing MLE’s for $\lambda_j$ and the $p_{wj}$’s, using this model\(^1\). If there are ways to simplify your algorithm so that you don’t actually have to call optim() or some other optimizer in R, show how the simplification works.

(b) Implement your algorithm (simplified form, if you came up with something) on the 1972 data only.

---

\(^1\)So, forget about priors and just obtain MLE’s. Also, for this exercise, you do NOT have to calculate standard errors.