1 Spring 2013: Problem 3

3.1a) Statements 1, 3, and 4 should be marked (TRUE).

Explanations:
Statement 1 is TRUE because it is the optimality condition: we are just saying that the gradient of $J(\theta)$ at $\hat{\theta}$ is zero.

Statement 2 is FALSE because the greater than sign should be a less than sign: the optimal $\hat{\theta}$ minimizes $J(\theta)$, not maximizes it!

Statement 3 is TRUE because increasing λ means we enforce more regularization.

Statement 4 is TRUE because we can always add frivolous features without increasing the training error for the optimal $\hat{\theta}$ (for example, we could set the coefficents of $\hat{\theta}$ corresponding to those features to 0). However, note that it may take longer for our learning algorithm to *find* this $\hat{\theta}$!

3.1b) A good classifier here would be $y = 1$ iff $\hat{\theta} \cdot \phi(x) \geq 0.5$.

Explanation: What you *don't* want to do is threshold at 0, i.e. $y = 1$ iff $\hat{\theta} \cdot \phi(x) \geq 0$. This is because the target ratings (training labels) are 0 or 1, so the regression function that we learn will tend to predict values that are between 0 and 1. This means we should use the midpoint value 0.5 as our threshold.

3.1c) The predictions will tend towards 0.

Explanation: If we increase λ, then $\|\hat{\theta}\|$ will decrease. As a result, the regression function values $\hat{\theta} \cdot \phi(x)$ will tend towards zero. Given the decision rule above, the predictions are going to become biased towards $y = 0$.