Alternative project proposals for Bayesian theory 2010:

1. Implement the Bayesian predictive 'spam' classifier derived in the lecture materials and compare numerically its characteristics to the earlier derived maximum likelihood classifier when the amount of training data is a) small, b) large. Investigate the effect of the number of variables (#distinct words) included in the model by doing the analysis with 5, 10 and 15 variables. Report the distributions of the posterior classification probabilities and the classification accuracies for both methods.

2. Use approximate Bayesian Model Averaging (BMA) to represent model uncertainty in logistic regression concerning which predictor variables among candidates should be included in the model. This can be done with the BMA package for R software:
http://www2.research.att.com/~volinsky/bma.html and
http://cran.r-project.org/web/packages/BMA/index.html

You can choose from the following two options:
a) simulate a dataset with 10 candidate predictors and 50 observations per response group (binary response). Choose the generating model such that 3 independent predictors each have a small to moderate effect on the log-odds for the response variable, while the remaining 7 predictor candidates are all independent of the response. You can use independent distributions of your choice for these variables, but give them distinct shapes/locations.
b) Use a suitable dataset from the literature to demonstrate how the BMA approach works in this context. For instance, various packages in R contain such datasets.

3. Assignments T14 & T15 were concerned with model selection for contingency table data related to marginal independence and/or conditional independence of the observed variables. Extend the investigation of model posterior probabilities by examining how they behave in the same setting of 3 binary variables as a function of the Dirichlet hyperparameter λ (assume all cells in the contingency table have the same value on λ_i) and the total number of observations n in the table. Choose two different models from the set of 8 possible models discussed in the assignment T15 and simulate data under them for a contingency table. By varying the number of observations and the Dirichlet hyperparameter, you can conclude how the two factors affect Bayesian inference in this context.

4. Model comparison procedures may behave erratically when none of the proposed models provides a reasonable approximation to the distribution of the observed data. Consider a situation where two datasets arise as follows. In the dataset A, 50 values of parameter θ are first drawn from a Beta(α,β) distribution, and conditional on each value, X is sampled from the corresponding Bernoulli(θ) distribution. The dataset B is generated equivalently, except that θ is drawn from Beta(μ,ψ) distribution. Use Bayes factors and posterior model probabilities to compare the following two standard Binomial(n,θ) models for these data: M1 one stating that the θ is the same for datasets A and B, and the other model M2 stating that θ is different in the two datasets. You should use the two alternative priors Beta(1/2,1/2) and Beta(1,1) for the success probability in the Binomial models. Notice that under M2 you need to assign a Beta prior to both Binomial models (datasets A & B). Examine the behavior of the Bayes factors and posterior model probabilities as a function of the number of observations available, such that these model comparison quantities are successively calculated from the *i* first observations in both datasets, *i* = 1,…,50. Plot the two functions over the range [1,50]. Examine the effect of varying the values in pairs (α,β) and (μ,ψ), such that they are small (<2) or moderate (>10). Consider the three cases where the two Beta distributions are equal, have the same expectation and different expectations. Notice that you here wrongly assume standard Binomial models for your datasets, which in fact are overdispersed as the underlying Bernoulli probabilities vary over the observations.

5. Examine in the context of models in project #3 above, the differences between asymptotic approximations of marginal likelihood, based on the two variants of Laplace approximation (formulae 4 & 5) and Schwarz approximation above formula (9) in Kass and Raftery (1995): http://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf
Derive the values of these approximations for the models considered in project #3 above and compare behavior of them and the corresponding approximate posterior model probabilities as a function of the total number of observations in the table. Pay in particular attention to how the approximations behave for small sample sizes. Simulate the data as described in project #3.