**Cogs 109: Modeling and Data Analysis**

Midterm study guide and topic list

For the midterm exam on Tuesday 10/30 (in class):

- Calculators allowed but no smartphones. Calculators will not be necessary — calculations will be relatively simple.
- There will not be any coding exercises. However, you may be asked to describe in words how you would perform a particular analysis.
- Content will be similar to HW1 through HW4. Anything covered in class up through Wednesday 10/24 is fair game. The corresponding textbook chapters are Ch. 1 through 5, although note that there are some topics in the book that we did not cover in class -- these will not be on the exam.
- If you understand the solution to all of the HWs and all of the lecture material, you should be in good shape.
- For additional practice, you could try to work out the exercises from the textbook that were not assigned on the HW. The following exercises are particularly relevant:
    - Chapter 2, exercises 2-6
    - Chapter 3, exercise 1
    - Ch. 4, ex. 1, 2, 5, 8, 9
    - Ch. 5, ex. 3, 4
- **[Updated] 1 page (single sided) of handwritten notes are allowed.**

Topics and keywords:

1. Types of models and modeling goals
    a. Prediction and Inference
    b. Parametric and Non-parametric models
    c. Supervised and Unsupervised learning
    d. Regression and Classification
    e. Training and Testing Error
    f. Flexible/Complex and Inflexible/Simple models
2. Given a model such as $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, identify the predictors, response, parameters, and noise
3. Bias, variance, irreducible error, total error
4. Tradeoff between flexibility and interpretability of models
5. Training vs. testing error
6. Overfitting
7. Bayes' theorem: $P(Y|X) \propto P(X|Y)P(Y)$. Identify the prior, posterior and likelihood terms.
8. Regression
    a. Leverage
    b. Standard deviation, standard error
    c. Confidence intervals
    d. Null hypothesis, p-value, t-statistic
    e. Residual
    f. Multiple regression: Explain how and why the result of a multiple regression may be different depending on which predictors are included
    g. Correlation, $R^2$
    h. Dummy variables: Explain how to interpret the coefficients of a regression model with dummy variables representing categorical predictors
9. Classification
    a. K-nearest neighbors
    b. Logistic regression
    c. LDA: Linear discriminant analysis

    d. Odds ratio, log-odds, logistic function
    e. Bayesian classifier
    f. Prior probability, posterior probability, data likelihood
    g. Decision boundary, decision threshold
    h. Confusion matrix
        i. Errors: False positive, False negatives
        ii. Sensitivity, specificity
    i. ROC analysis, ROC curve
10. Resampling and cross-validation
    a. Validation set
    b. Using cross-validation to compare models (e.g. linear vs. quadratic regression)
    c. LOOCV, k-fold CV
    d. Bootstrap resampling