Midterm 2 information and review
– Math 313 Statistics for Data Science

Guangliang Chen

Associate Professor

cheng@hope.edu

Hope College, Fall 2023

Presentation Overview

1. General information

2. Review of the material

3. Practice problems

General information

- **Time and location**: November 22, Wednesday, in class (50 minutes).
- **Rough range of material to be tested**: Section 2.2.3 and Chapters 4 of the textbook
- **Study resources**: This study guide, lecture slides, and practice problems.
- **Cheat sheet**: One page of notes may be used (you can write on the back of the notes for midterm 1, or on another sheet of paper). No other resources will be allowed during the exam.
- **Electronic device policy**: A scientific calculator (but not cell phones, iPads, laptops, etc.) may also be used, but it is not necessary.

General information (cont'd)

- **Type of questions**: Roughly, there are 3 types of questions:
  - conceptual questions that test your basic knowledge of the concepts and methods (for example, multiple-choiced or short-answer questions about various classifiers - kNN, logistic regression, LDA/QDA, Naive Bayes),
  - computational questions that are very analogous to lecture examples on LR, LDA and naive Bayes with a single predictor
- **Important reminder**: Show your work on all exam questions (except the multiple-choiced ones) to receive full credit
  - An incorrect answer without any work will receive zero points
  - An incorrect answer with some correct work will receive partial credit
- **Review session**: We will reserve next Monday's class for going through the practice questions and answering any questions you might have.
- **Additional office hours**: 9-11:30am, November 22, Wednesday (I am also available on the preceding day by appointment).

- **Classification**
  - What is a classifier, training error rate and test error rate
  - What are Bayes classifiers: posterior probability, Bayes decision boundary and Bayes error rate, kNN classification

Chapter 4 review

- **Logistic regression (LR)**:
  - Problem setup with the logistic loss
  - The statistical interpretation of LR, as well as the likelihood function
  - How to classify new observations
  - Multiclass extensions (one vs one, one vs rest, multinomial)
  - LR as a generalized linear model (link function, distribution)
- **Discriminant analysis**:
  - The statistical perspective of classification, model-based Bayes classification, prior and posterior probabilities
  - LDA in 1 dimension, can work out the math
  - LDA in higher dimensions (multivariate Gaussians with shared covariance matrix)
  - QDA in any number of dimensions, and how it differs from LDA
  - Naive Bayes (Gaussian NB and Bernoulli NB)
  - Poisson regression

See Chapter 4 lecture slides (there are several practice problems at the end of each topic):

- Show that for the sigmoid function $g(z) = \frac{1}{1+e^{-z}}$, we must have $g'(z) = g(z)(1 - g(z))$ for all $z \in \mathbb{R}$. Where is $g'(z)$ largest?
  Answer: $z = 0$

- It was mentioned that logistic regression uses the logit link function (which leads to the sigmoid function):

$$\log \frac{p}{1-p} = \boldsymbol{\theta} \cdot \mathbf{x} \quad \longrightarrow \quad p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta} \cdot \mathbf{x}}}$$

  Other choices of the link function include the following:
  - **Cauchit (inverse Cauchy)**: $\tan(\pi(p - 0.5))$
  - **Probit**: $\Phi^{-1}(p)$, where Φ is the cdf of standard normal.
  - **Complementary log-log**: $\log(-\log(1-p))$
  - **Negative log-log**: $-\log(-\log(p))$

  Derive a formula for p (in terms of $\mathbf{x}$ and $\boldsymbol{\theta}$) by using the Cauchit link function. Answer: $p = \frac{1}{\pi}\arctan(\boldsymbol{\theta} \cdot \mathbf{x}) + \frac{1}{2}$

- We apply Gaussian Naive Bayes (GNB) to solve a binary classification problem with two predictors x_1, x_2. Suppose that
    - Within class 1, we fit a $N(0, 2^2)$ distribution over the values of x_1 and a $N(0, 2^2)$ distribution over x_2;
    - Within class 2, we fit a $N(4, 2^2)$ distribution over the values of x_1 and a $N(0, 1^2)$ distribution over x_2;

  The prior probabilities of the two classes are $\pi_1 = \frac{2}{3}$ and $\pi_2 = \frac{1}{3}$. What is the decision boundary of the GNB classifier?
  Answer: $x_1 = \frac{3}{8}x_2^2 + 2$

- Explain in what context (in terms of response) should we perform each of the following regressions:
  - Linear regression: $y \mid x \sim N(\mu, \sigma^2)$ with $\mu = \beta_0 + \beta_1 x$
  - Logistic regression: $y \mid x \sim \mathrm{Bernoulli}(p)$ with $p = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$
  - Poisson regression: $y \mid x \sim \mathrm{Poisson}(\lambda)$ with $\lambda = e^{\beta_0+\beta_1 x}$

  Write down the corresponding likelihood functions in the case of a single predictor (when given a sample of size n: $(x_1, y_1), \ldots, (x_n, y_n)$):
  - Linear regression: $\ell(\beta_0, \beta_1 \mid \{(x_i, y_i)\}_{i=1}^n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{-(y_i-\mu_i)^2}{2\sigma^2}}$, where $\mu_i = \beta_0 + \beta_1 x_i$
  - Logistic regression: $\ell(\beta_0, \beta_1 \mid \{(x_i, y_i)\}_{i=1}^n) = \prod_{i=1}^n p_i^{y_i}(1-p_i)^{1-y_i}$, where $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$
  - Poisson regression: $\ell(\beta_0, \beta_1 \mid \{(x_i, y_i)\}_{i=1}^n) = \prod_{i=1}^n \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i}$, where $\lambda_i = e^{\beta_0+\beta_1 x_i}$

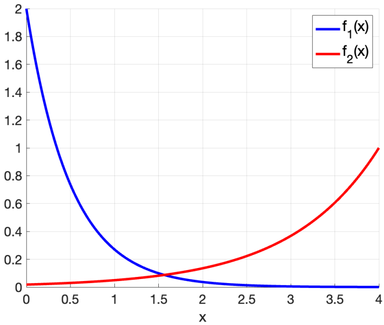
Additional practice problems

1. Which of the following classifiers is not based on the posterior probability?
   (A) kNN
   (B) Logistic regression
   (C) LDA/QDA
   (D) Naive Bayes
   (E) None of the above (they all use the posterior probability to make predictions)

2. A mixture distribution with two components in 1 dimension has the following density function

$$f(x) = 0.6 f_1(x) + 0.4 f_2(x)$$

where $f_1(x) = 2e^{-2x},\ x > 0$ and $f_2(x) = e^{x-4},\ x < 4$. One can construct a Bayes classifier based on this model similarly to 1D LDA. What is the decision boundary? Answer: $x = 1.6995$



3. The values of two predictors (when restricted to class 1) in a classification problem are shown below. The first predictor is numerical while the second one is binary. It is possible to mix Gaussian NB and Bernoulli NB in this case to model each class (under the independence assumption). Fit a Gaussian distribution to x_1 and a Bernoulli distribution over x_2, both within class 1. What is the likelihood of a test point $(0.8, 1)$ relative to class 1?
Answer: $N(\mu = 0.8, \sigma^2 = 0.37)$ and Bernoulli($p = \frac{2}{3}$).
Likelihood of test point relative to class 1 = 0.4372

	x_1	x_2
Class 1	0.5	1
	1.5	0
	0.4	1
Class 2		

4. How does each of the following classifiers handle the multiclass setting (i.e., 3 or more classes)?
   - kNN
   - Logistic regression
   - LDA/QDA
   - Naive Bayes

Multinomial logistic regression is a neural network