APPLICATION FOR NEW COURSE

1. Submitted by College of ___________ Public Health _____________________________________________________________________________ Date ___________ 20 October 2004 _____________________________________________________________________________
Department/Division offering course ________________ Biostatistics _____________________________________________________________________________

2. Proposed designation and Bulletin description of this course

   a. Prefix and Number ________________ CPH 636 ________________ b. Title* ________________ Data Mining in Public Health _____________________________________________________________________________
   *NOTE: If the title is longer than 24 characters (including spaces), write A sensible title (not exceeding 24 characters) for use on transcripts ________________ Data Mining _____________________________________________________________________________

   c. Lecture/Discussion hours per week ________________ d. Laboratory hours per week ________________

   e. Studio hours per week ________________ f. Credits ________________

   g. Course description

   This course concerns statistical techniques for and practical issues associated with the exploration of large public health data sets, the development of models from such data sets, and the effective communication of one's findings.

   h. Prerequisites (if any)

   required: STA 570 or STA 580 or consent of instructor

   recommended but not required: CPH 535

   i. May be repeated to a maximum of ________________ (if applicable)

4. To be cross-listed as

   Prefix and Number ________________ _____________________________________________________________________________ Signature, Chairman, cross-listing department _____________________________________________________________________________

5. Effective Date ________________ Spring 2006 ________________ (semester and year)

6. Course to be offered

   ☐ Fall  ☑ Spring  ☐ Summer

7. Will the course be offered each year?

   ☑ Yes  ☐ No (Explain if not annually)

8. Why is this course needed?

   Large public health data sets, if examined effectively, can yield valuable insights to public health practitioners as they seek to monitor the well-being of a population, identify opportunities for interventions, and evaluate the results of policies.

9. a. By whom will the course be taught? ____________________________ Marta Mendiondo and/or Richard Charnigo _____________________________________________________________________________

   b. Are facilities for teaching the course now available?

   ☑ Yes  ☐ No

   If not, what plans have been made for providing them?

__________________________________________________________________________

__________________________________________________________________________
APPLICATION FOR NEW COURSE

10. What enrollment may be reasonably anticipated? 10-15 students

11. Will this course serve students in the Department primarily?
Yes □ No □

Will it be of service to a significant number of students outside the Department?
Yes □ No □

Public health students who want the skills to gain insights from large data sets will benefit from this course, regardless of their concentration area. Although the course will be geared toward a public health audience, students from medicine, nursing, and other disciplines may also be interested in enrolling; many of the statistical techniques and practical issues are general in their applicability.

Will the course serve as a University Studies Program course?
Yes □ No □

If yes, under what Area?

12. Check the category most applicable to this course

traditional; offered in corresponding departments elsewhere;
relatively new, now being widely established
not yet to be found in many (or any) other universities

13. Is this course applicable to the requirements for at least one degree or certificate at the University of Kentucky?
Yes □ No □

14. Is this course part of a proposed new program?
Yes □ No □
If yes, which?

15. Will adding this course change the degree requirements in one or more programs?*
Yes □ No □
If yes, explain the change(s) below

Biostatistics students will be able to take CPH 636 as a concentration-area selective.

Other departments may see fit to allow their students to take CPH 636 as a concentration-area selective.

16. Attach a list of the major teaching objectives of the proposed course and outline and/or reference list to be used.

17. If the course is a 100-200 level course, please submit evidence (e.g., correspondence) that the Community College System has been consulted.

18. If the course is 400G or 500 level, include syllabi or course statement showing differentiation for undergraduate and graduate students in assignments, grading criteria, and grading scales.

19. Within the Department, who should be contacted for further information about the proposed course?

Name Marta Mendiondo
Phone Extension 859 257-1412 x274
APPLICATION FOR NEW COURSE

Signatures of Approval:

[Signatures]

Date

[Signatures]

Date

Date of Notice to the Faculty

Date

Date

Date

Date of Notice to University Senate

*If applicable, as provided by the Rules of the University Senate

ACTION OTHER THAN APPROVAL

Rev 3/04
*NOTE: Approval of this course will constitute approval of the program change unless other program modifications are proposed.
CPH 636: Data Mining in Public Health

Course Description: This course familiarizes students with statistical techniques for and practical issues associated with the exploration of large public health data sets. Emphases are placed on the development of models from such data sets and on the effective communication of one’s findings.

Course Structure: 3 credit hours (2 hours of lecture, 2 hours of laboratory)

Prerequisite: STA 570 (Basic Statistical Analysis) or STA 580 (Biostatistics I) or consent of instructor; CPH 535 (Databases and SAS Programming) recommended but not required

Initial Offering: Spring 2006

Instructors: Marta Mendiondo and/or Richard Charnigo

Course-Specific Objectives: Upon completion of CPH 636, students will be able to do the following:

1. Articulate the challenges associated with the acquisition and analysis of large public health data sets;
2. Judiciously apply linear regression methodology to problems in public health;
3. Judiciously apply linear classification methodology, including logistic regression, to problems in public health;
4. Employ classification and regression trees to analyze large public health data sets;
5. Employ neural networks and nearest neighbor methods to analyze large public health data sets;
6. Articulate the strengths and weaknesses of various supervised learning techniques; and,
7. Apply unsupervised learning techniques to problems in public health.

References: Material will be taken from:

1. [required for students] lecture notes prepared specifically for CPH 636;
2. [recommended for students] Fernandez, Data Mining Using SAS Applications, Chapman and Hall (2003);
3. [optional for students] Stroup and Teutsch, Statistics in Public Health, Oxford (1998);
4. [not for students, relevant material will be presented by instructors at the students’ level] Ripley, Pattern Recognition and Neural Networks, Cambridge (1996); and,
5. [not for students, relevant material will be presented by instructors at the students’ level] Hastie, Tibshirani, and Friedman, The Elements of Statistical Learning, Springer (2001).
Overview of Topics:

1. Motivating examples from public health; sources of public health data; statistical modeling as a remedy to the curse of dimensionality; training data, testing data, and model assessment; organizing large data sets; univariate and bivariate exploratory analyses.

2. Introduction to supervised learning; simple linear regression; multiple linear regression; interpretations and inferences in linear regression; variable selection and shrinkage in linear regression; criteria for model assessment.

3. Regression problems versus classification problems; Bayes’ Theorem and the receiver operator curve; linear discriminant analysis; quadratic discriminant analysis; simple logistic regression; multiple logistic regression; interpretations and inferences in logistic regression; criteria for model assessment.

4. Regression trees; classification trees; handling missing information when using regression and classification trees; general guidelines for handling missing information.


6. Criteria for comparing supervised learning techniques, including interpretability, predictive power, computational feasibility, and robustness to outliers; comparative assessments of linear techniques, trees, neural networks, and nearest neighbor methods.

7. Introduction to unsupervised learning; principal components; factor analysis; K-means clustering; hierarchical clustering.

Assessment: There will be an “open-book” practical examination (30%), a final project entailing a short oral presentation (15%) and written report (15%), and regular laboratory/homework activities (40%).