**PH.140.644.01**
**STATISTICAL MACHINE LEARNING: METHODS, THEORY, AND APPLICATIONS**
**JOHNS HOPKINS BLOOMBERG SCHOOL OF PUBLIC HEALTH**
**AY 2026–2027, TERM 1 (FALL 2026)**

**Instructor:** Yiqun Chen, PhD
**E-mail:** yiqunc@jhu.edu
**Office Hours:** Tuesdays, 3:00–3:50 PM, Wolfe 2015, and by appointment

**Teaching Assistant:** Junwei (Ivy) Sun
**E-mail:** jsun140@jh.edu
**Office Hours:** TBD (announced on CoursePlus)

**Communication:** Use email for logistics and other non-course-material matters. Course-material questions and discussion should be posted on CoursePlus.

**Prerequisites:** Linear algebra, intermediate statistics, and basic R programming

**Course Description:**

This course introduces the statistical and computational foundations of modern machine learning for public health and biomedical research. Topics include regression, classification, resampling methods, model selection and regularization, flexible and nonlinear modeling, tree-based methods and ensembles, unsupervised learning, and an introduction to deep learning and modern AI. The course emphasizes implementation, reproducible workflows, careful model evaluation, scientific interpretation, and responsible use of AI tools.

**Learning Objectives:**

Upon completion of this course, students will be able to:

- Identify appropriate machine learning methods for different data structures and public health research questions
- Translate scientific questions into supervised, unsupervised, and predictive learning problems
- Fit, tune, and evaluate models using reproducible workflows
- Explain the bias-variance tradeoff and its implications for model complexity and prediction
- Apply cross-validation and held-out test sets for model assessment
- Interpret regression, classification, regularization, tree-based, ensemble, and unsupervised learning results
- Recognize common threats to validity, including overfitting, data leakage, poor calibration, and inappropriate performance metrics
- Critically evaluate modern AI and large language model applications in biomedical and public health settings
- Communicate analyses clearly, including limitations and uncertainty

**Assessment Methods:**

- **Homework Assignments (40%):** Four assignments, each graded out of **100 points** and counting for 10% of the final grade. All four count; flex days are the buffer for a bad week.
- **In-Class Quizzes (50%):** Six short individual quizzes given at the start of Thursday class. Each is graded out of **100 points** and designed to take about 15 minutes. The **best five scores** count at 10% each; the lowest is dropped automatically.
- **Participation (10%):** Engagement in live activities, CoursePlus discussions, and brief reflection prompts.

| Component | Count | Each | Total |
|---|---|---|---|
| Homework | 4 assignments, all count | 10% | 40% |
| In-class quizzes | 6 given, best 5 count | 10% | 50% |
| Participation | — | — | 10% |
| **Total** | | | **100%** |

Homework assesses applied fluency and communication. Short in-class quizzes verify individual conceptual understanding and reasoning about model output.

**Structure of Each Assessment:**

Every homework and every quiz is scored out of 100 points, with point values printed next to each question. Quizzes are closed-book, cover the preceding material, and require no computation beyond simple arithmetic.

**Use of AI Tools:**

There is currently no official, school-wide policy governing the use of generative AI in courses at the Bloomberg School. In the absence of one, this course follows the University's Guidelines for Responsible Use of AI (`https://it.johnshopkins.edu/ai/guidelines-for-responsible-use-of-ai/`), together with the expectations below.

- **Use AI critically and responsibly.** These tools are useful for studying, debugging, syntax help, and brainstorming. They also produce confident errors, fabricated citations, and code that runs but answers the wrong question. Check what they give you.
- **You stand behind everything you submit.** You are responsible for every statement, number, figure, and line of code in your submitted work, whether or not AI was involved. "The model wrote it" is not a defense of an incorrect or unsupported claim.
- **Pasting AI output directly is strongly discouraged.** Copying generated text or code without working through it is an ineffective way to learn the material, and it is precisely the skill the in-class quizzes are designed to test.
- **Disclose substantive assistance.** If AI materially shaped a homework answer, note briefly where and how. A one-line note is enough.

- **No AI during in-class quizzes.**

These expectations are in addition to, and do not replace, any applicable School and University regulations on academic integrity, plagiarism, and the use of AI. Where those regulations are stricter, they govern.

**Required Textbook:**

*An Introduction to Statistical Learning, with Applications in R (2nd Edition)*, by James, Witten, Hastie, and Tibshirani. Free at `http://www.statlearning.com`.

**Supplementary Resources:**

- *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman, for advanced reference
- Course-specific code, datasets, and worked examples posted on CoursePlus
- Selected articles and public health case studies on responsible machine learning and AI

**Computing Environment:**

Course examples, reference code, and troubleshooting support focus on R and RStudio. Python is also supported. Students using another language are responsible for translating examples and ensuring reproducible submissions.

**Course Policies:**

**Late Submission Policy:**

- Each student receives 4 flex days to use across the four homework assignments, however they choose, with no explanation required
- After flex days are exhausted, late homework receives a 10% penalty per day
- Homework more than 5 days late is not accepted unless prior approval has been granted
- Because there are only four assignments, no homework score is dropped; the flex days exist for this reason
- Quizzes are not normally made up because the lowest quiz score is dropped; documented emergencies are handled under BSPH policy

**Collaboration and Academic Integrity:**

- Collaboration is encouraged for conceptual discussion and study
- Submitted work must reflect each student's own understanding
- Students may discuss strategy and pseudocode, but may not copy or share full solutions
- External resources, including AI tools, must be cited or disclosed when they materially shape submitted work

**Accessibility:** Students with documented disabilities should contact Student Disability Services. Course materials will be made available in accessible formats when possible.

**Tentative Course Schedule:**

Class meets Tuesdays and Thursdays, 1:30–2:50 PM, in Wolfe 2015. Quizzes are given at the start of Thursday class. Homework is due at 11:59 PM Eastern Time through CoursePlus, on Friday except HW3 (Sunday). All dates are tentative; the CoursePlus calendar is authoritative.

| Wk | Date | Session | Due |
|---|---|---|---|
| 1 | Tue Sep 1 | Class organization; overview of statistical learning | |
| | Thu Sep 3 | Overview of statistical learning and modern data science | |
| 2 | Tue Sep 8 | Regression: recap and new ideas | |
| | Thu Sep 10 | Classification | **Quiz 1** |
| | Fri Sep 11 | | **HW1** |
| 3 | Tue Sep 15 | Classification and evaluation | |
| | Thu Sep 17 | Regularization: ridge regression and the lasso | **Quiz 2** |
| 4 | Tue Sep 22 | Cross-validation | |
| | Thu Sep 24 | Model selection: choosing tuning parameters | **Quiz 3** |
| | Fri Sep 25 | | **HW2** |
| 5 | Tue Sep 29 | Non-linear methods I | |
| | Thu Oct 1 | Non-linear methods II | **Quiz 4** |
| 6 | Tue Oct 6 | Deep learning | |
| | Thu Oct 8 | Moving beyond tabular data | **Quiz 5** |
| | Sun Oct 11 | | **HW3** |
| 7 | Tue Oct 13 | Unsupervised learning: dimension reduction | |
| | Thu Oct 15 | Unsupervised learning: clustering | |
| 8 | Tue Oct 20 | Large language models | |
| | Thu Oct 22 | Effective applications of ML/AI methods | **Quiz 6** |
| | Fri Oct 23 | | **HW4** |

**Homework Assignments:** *Each assignment is worth 100 points and 10% of the grade.*

- **HW1** — due Fri Sep 11. *Statistical learning foundations and linear regression.* Covers Week 1 and the Week 2 regression session.
- **HW2** — due Fri Sep 25. *Classification and evaluation.* Covers the Week 2 classification session and the Week 3 Tuesday session.
- **HW3** — due Sun Oct 11. *Model selection, regularization, and non-linear methods.* Covers the Week 3 Thursday session and Weeks 4 and 5.

- **HW4** — due Fri Oct 23. *Deep learning and text as data.* Covers Week 6, with one conceptual item from Week 7.

**Quizzes:** *Each quiz is worth 100 points; the best five of six count.*

- **Quiz 1** (Thu Sep 10) — Statistical learning foundations
- **Quiz 2** (Thu Sep 17) — Regression and classification
- **Quiz 3** (Thu Sep 24) — Classification, model flexibility, and KNN
- **Quiz 4** (Thu Oct 1) — Regularization and cross-validation
- **Quiz 5** (Thu Oct 8) — Non-linear methods, trees, and ensembles
- **Quiz 6** (Thu Oct 22) — Deep learning and non-tabular data