Proposal for a New Course

Department of Computer Science and Engineering
Indian Institute of Technology, Kanpur

Course number: IS201
Course title: Probability and Statistics for AI-ML
Course prerequisites: MTH111 and MTH112 (waived for PG students).
Course credits: [10] (3-0-1-0)
Course duration: Full semester
Course type: DC (for IS UG students)
Proposing instructor(s): Nitin Saxena
Other faculty members interested in teaching the course: Sayak Ray Chowdhury
Other departments interested in the proposed course: None

Course description:
a. **Objectives:** This is a probability and statistics course tailored for applications in CSE, AI, and Machine Learning (ML). It will bridge theoretical foundations with practical application, focusing on quantifying uncertainty, modeling data distributions, and inferential techniques. Key topics include probability distributions, Bayesian inference, maximum likelihood estimation (MLE), hypothesis testing, Markov Chain, and analysis to enable model evaluation and understanding of algorithm behavior.
b. **Logistics**: The course will serve as a DC for IS UG students (BT, double major, digital IS minor), and as a DE for PG students (MT, MS, PhD) of the CSE and IS departments and CSE UG students (BT, double major, ML minor). The course will be offered once every academic year.
c. **Content:** There will be an equivalent of 26 lectures of 75 minutes each and 12 labs of 1 hour each (which can also serve as a tutorial for query resolution and quizzes).
d. **Evaluation:** Evaluation will use a combination of lab exercises, take-home assignments and projects, and traditional sit-down quizzes and exams.

Weekly breakup of content (26 lectures x 75mins):

<u>Week 1:</u> *Concepts:* Probability definition. Set operations. *Applications:* Paradoxes. Examples. Handling Uncertainty.

<u>Week 2:</u> *Concepts:* Derangements. Sigma algebra. Conditional probability. *Applications:* Infinite sample space. Examples. Statistical Inference & Estimation: Maximum Likelihood Estimation (MLE) and Maximum a posteriori (MAP) estimation to fit models.

Week 3: *Concepts:* Dependence of events. Bayes theorem. Random variable (RV). *Applications:* Hypothesis testing. Monty Hall fallacy. N-gram algorithm. Naive Bayes classifier algorithm. Spam filtering algorithm. Birthday Paradox.

Week 4: *Concepts:* RV independence. Descriptive statistics (Expectation, Standard deviation, Covariance, Pearson's correlation coefficient). Conditional distribution. Conditional expectation and its linearity. *Applications:* Examples. Lloyd's k-means clustering algorithm. Data Analysis & Statistics. Precision, Recall, F1-score, and P-R Curve.

Week 5: *Concepts:* Information Theory (Entropy, Mutual information, Joint entropy, Information gain). Important RVs (Bernoulli, Binomial, Geometric, Negative binomial, Exponential, Normal/Gaussian). Probability density function. *Applications:* Decision tree learning *(*Quinlan's ID3) algorithm. Stating the Central Limit Theorem: physical interpretation of the expectation.

Week 6: *Concepts:* Poisson RV. Concentration inequalities (Markov, Chebyshev, Chernoff). Weak-linearity of variance. Weak-law of large numbers. *Applications:* Buffon's needle problem. Equality testing algorithm.

Week 7: *Concepts:* Chernoff's proof. K-wise independence. Gaussian mixture model (GMM). *Applications:* Boosting the success probability of algorithms. Multi-server load balancing algorithm. EM algorithm to learn GMM.

Week 8: *Concepts:* Linear regression. Logistic regression. Stochastic process. Markov Chain (MC). Evolution of a MC. *Applications:* Regression algorithms. Random walk (higher-dimension grid, and graphs).

Week 9: *Concepts:* Regular, homogeneous MC. Stationary distribution. Ergodic MC. Martingale. *Applications:* Page rank (Internet-search) algorithm. Cell genetics; Dominating genes.

Week 10: *Concepts:* Uniform sampling– k numbers, permutation. Biased-coin tossing. Monte Carlo Markov Chain (MCMC) sampling. *Applications:* Sampling algorithms. Metropolis-Hastings algorithm for MCMC.

Week 11: *Concepts:* Hash functions. Probabilistic method. *Applications:* Implementation of a hash over GF(2). The existence of the Ramsey number.

Week 12: *Concepts:* Probabilistic method. *Applications:* The existence proofs (and randomized algorithms) for large cuts, sum-free subsets, extremal set families, and superconcentrators.

Week 13: *Concepts:* The data stream model. *Applications:* Streaming algorithms for– counting, moments, and heavy-hitter detection.

Short summary for inclusion in the Courses of Study booklet: This is a probability and statistics course tailored for applications in CSE, AI, and Machine Learning (ML). It will bridge theoretical foundations with practical application, focusing on quantifying uncertainty, modeling data distributions, and inferential techniques. The presented concepts enable model evaluation and understanding of algorithm behavior.

Textbook: <<Uses assorted online sources for various topics>>

- Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.
- Probability, Random Variables, and Stochastic Processes. Authors: Athanasios Papoulis
- Introduction to Mathematical Statistics. Authors: Robert V. Hogg, J.W. McKean, and Allen T. Craig

Course proposer: Nitin Saxena.

Date: 07-Feb-2026

The course is approved/not approved

Chairperson, SPGC
Date: