UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Astrostatistics and Machine Learning

2526-1-F5803Q020

Obiettivi

L'uso della statistica è onnipresente in astronomia e astrofisica. I progressi moderni sono resi possibili dall'applicazione di strumenti sempre più sofisticati, spesso noti come "data mining", "machine learning" e "intelligenza artificiale". Questo corso fornisce un'introduzione ad alcune di queste tecniche statistiche in modo altamente pratico, combinando derivazioni formali con applicazioni computazionali dirette. Sebbene gli esempi saranno presi quasi esclusivamente dal campo dell'astronomia, il corso è adatto a tutti gli studenti di Fisica interessati al machine learning.

Gli studenti acquisiranno competenze in inferenza statistica, analisi dei dati e calcolo avanzato, con esperienza pratica nell'applicazione di tecniche di machine learning a dati (astro)fisici.

Contenuti sintetici

- Probabilità e statistica.
- Inferenza frequentista e bayesiana.
- Machine learning.

Programma esteso

1. Introduction

- Data mining and machine learning.
- Supervised and unsupervised learning.
- Python setup.

- Version control with git.

2. Probability and Statistics

- Probability.
- Bayes' theorem.
- Random variables.
- Monte Carlo integration.
- Descriptive statistics.
- Common distributions.
- Central limit theorem.
- Multivariate pdfs.
- Correlation coefficients.
- Sampling from arbitrary pdfs.

3. Frequentist Statistical Inference

- Frequentist vs Bayesian inference.
- Maximum likelihood estimation.
- Omoscedastic Gaussian data, Heteroscedastic Gaussian data, non Gaussian data.
- Maximum likelihood fit.
- Role of outliers.
- Goodness of fit.
- Model comparison.
- Gaussian mixtures.
- Boostrap and jackknife.
- Hypotesis testing.
- Comparing distributions, KS test.

4. Bayesian Statistical Inference

- The Bayesian approach to statistics.
- Prior distributions.
- Credible regions.
- Parameter estimation examples
- Marginalization.
- Model comparison: odds ratio.
- Approximate model comparison.
- Monte Carlo methods.
- Markov chains.
- Burn-in.
- Metropolis-Hastings algorithm.
- MCMC diagnostics: burn-in, autocorrelation lenght, traceplots.
- Samplers in practice: emcee and PyMC3.
- Gibbs sampling.
- Conjugate priors.
- Evidence evaluation.
- Nested sampling.
- Samplers in practice: dynesty.

5. Clustering.

- K-fold cross validation.
- K-Means Clustering.
- Mean-shift Clustering.

- Correlation functions.

6. Dimensional Reduction

- Curse of dimensionality.
- Principal component analysis.
- Non-negative matrix factorization.
- Independent component analysis.
- Non-linear dimensional reduction.
- Locally linear embedding.
- Isometric mapping.
- t-distributed stochastic neighbor embedding.

7. Density estimation

- Histograms
- Kernel density estimation
- Nearest-Neighbor.
- Gaussian Mixtures.

8. Regression

- Linear regression.
- Polynomial regression.
- Basis function regression.
- Kernel regression.
- Over/under fitting.
- Cross validation.
- Learning curves.
- Regularization: Ridge, LASSO.
- Non-linear regression.
- Gaussian process regression.
- Total least squares.

9. Classification

- Generative vs discriminative classification.
- Receiver Operating Characteristic (ROC) curve.
- (Gaussian) Naive Bayes.
- Linear and quadratic discriminant analysis.
- GMM Bayes classification.
- K-nearest neighbor classifier.
- Logistic regression.
- Support vector machines.
- Decision trees.
- Bagging.
- Random forests.
- Boosting.

10. Deep learning

- Loss functions.
- Gradient descent, learning rate.
- Adaptive boosting.
- Neural networks.

- Backpropagation.
- Layers, neurons, activation functions, regularization schemes.
- Neural network in practice: TensorFlow, keras, and pytorch.
- Convolutional neural networks.
- Autoencoders.
- Generative adversarial networks.

Examples of astrophysical datasets we will be using:

- Galaxies and quasars from the Sloan Digital Sky Survey.
- Black-hole binaries from the Laser Interferometer Gravitational-Wave observatory (LIGO).
- Simulated supernova observations
- Gamma Ray bursts
- and more...

Prerequisiti

Nessun prerequisito formale. È **fortemente consigliata** una conoscenza preliminare del linguaggio di programmazione Python e familiarita' con la shell Unix (vedi sotto per alcune risorse di recupero).

Modalità didattica

Lezioni, 6 crediti.
Modalita': erogativa 50%, interattiva 50%.

Il data mining e il machine learning sono discipline computazionali. Non si impara a trattare i dati scientifici leggendo equazioni alla lavagna: bisogna mettersi alla prova (ed è questa la parte divertente!). Gli studenti sono tenuti a partecipare alle lezioni con un laptop o un dispositivo su cui sia possibile programmare (direi qualcosa di più grande di uno smartphone…). Ogni lezione combinerà spiegazioni teoriche con esercizi pratici e dimostrazioni. Questi aspetti sono il cuore del corso, quindi vi invitiamo a partecipare attivamente il più possibile.

Materiale didattico

Il **libro di testo principale** e'

"Statistics, Data Mining, and Machine Learning in Astronomy", Željko, Andrew, Jacob, and Gray. Princeton University Press, 2012.

È un libro eccellente che continuo a consultare nella mia ricerca. La biblioteca ne ha alcune copie; è possibile scaricare una versione digitale dal sito della biblioteca di Bicocca. Un aspetto che apprezzo particolarmente è che forniscono il codice dietro ogni singola figura: astroml.org/book_figures. Il modo migliore per affrontare questi argomenti è studiare l'introduzione nel libro, quindi prendere il codice e sperimentare. Assicuratevi di ottenere l'edizione aggiornata del libro (quella con la copertina nera, non arancione) perché tutti gli esempi sono stati aggiornati a Python 3.

Esistono molte altre ottime risorse in astrostatistica; ecco un elenco parziale.

- "Statistical Data Analysis", Cowan. Oxford Science Publications, 1997.
- "Data Analysis: A Bayesian Tutorial", Sivia and Skilling. Oxford Science Publications, 2006.
- "Bayesian Data Analysis", Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin. Chapman & Hall, 2013.
- "Python Data Science Handbook", VanderPlas. O'Reilly Media, 2016.
- "Practical Statistics for Astronomers", Wall and Jenkins. Cambridge University Press, 2003.
- "Bayesian Logical Data Analysis for the Physical Sciences", Gregory. Cambridge University Press, 2005.
- "Modern Statistical Methods For Astronomy" Feigelson and Babu. Cambridge University Press, 2012.
- "Information theory, inference, and learning algorithms" MacKay. Cambridge University Press, 2003.
- “Data analysis recipes”. Questi sono capitoli gratuiti di un libro non ancora completato da Hogg et al.:
    - "Choosing the binning for a histogram" [arXiv:0807.4820]
    - "Fitting a model to data [arXiv:1008.4686]
    - "Probability calculus for inference" [arXiv:1205.4446]
    - "Using Markov Chain Monte Carlo" [arXiv:1710.06068]
    - "Products of multivariate Gaussians in Bayesian inferences" [arXiv:2005.14199]
- "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow", Geron, O'Reilly Media, 2019.
- "Machine Learning for Physics and Astronomy", Acquaviva, Princeton University Press, 2023.

Faremo un ampio uso del linguaggio di programmazione Python. Se avete bisogno di rinfrescare le vostre competenze in Python, ecco alcune risorse di recupero e tutorial online. Una solida conoscenza di Python è essenziale nell’astrofisica moderna!

- "Scientific Computing with Python", D. Gerosa. Questo è un corso che tengo per la Scuola di Dottorato qui a Milano-Bicocca.
- "Lectures on scientific computing with Python", R. Johansson et al.
- Python Programming for Scientists", T. Robitaille et al.
- "Learning Scientific Programming with Python", Hill, Cambridge University Press, 2020. Supporting code: scipython.com.
- "Effective Computation in Physics", A. Scopatz, K. D. Huff, O'Reilly Media, 2015.

Periodo di erogazione dell'insegnamento

Secondo semestre.

Modalità di verifica del profitto e valutazione

L’esame consisterà in una prova orale. Durante il corso verrà assegnato un insieme di problemi computazionali, che gli studenti dovranno completare autonomamente e discutere durante l’esame, insieme a domande più generali sul materiale trattato.

L’esame valuterà:

- la capacità di estrarre informazioni statistiche dai dataset forniti;
- la conoscenza e la familiarità generale con gli argomenti trattati a lezione;
- la creatività e la padronanza nell’affrontare problemi concettuali e computazionali relativi alle tecniche studiate.

Tutte le lezioni, esercizi ed esami si svolgeranno in inglese.

Orario di ricevimento

In qualsiasi momento, potete contattarmi via email. Il mio ufficio è il numero 2007 nell’edificio U2.

Sustainable Development Goals

ISTRUZIONE DI QUALITÁ | IMPRESE, INNOVAZIONE E INFRASTRUTTURE