Introduction to Machine Learning for Economists: Lecture 1

Sonan Memon
Institute of Business Administration, Karachi

Dec 2020

Course Information

i. See course outline for full list of resources, readings and other information
ii. There will be four lectures, each lasting for 2 hours
iii. We will cover shrinkage estimators in Lecture 1, Bandit Problems in Lecture 2, Deep Neural Networks in Lecture 3 and Computational Linguistics or Text Analysis in Lecture 4
iv. Focus is on concepts, not computational implementation

Textbooks, Readings and Resources

i. See course outline for full list of resources, readings and other information
ii. Hastie, T., Tibshirani, R., & Friedman, J. (2017). The elements of statistical learning
iii. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: Springer.
iv. Course GitHub Page

Readings for Lecture 1

i. Alpaydin, Ethem (2010), Introduction to Machine Learning, 2nd edition, MIT Press, Chapter 1.

ii. Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Verlag, chapter 2.

iii. Hastie, T., Tibshirani, R., & Friedman, J. (2017). The elements of statistical learning, Chapter 3.4.

iv. Stachurski, John (2016), A Primer in Econometric Theory, MIT Press (Chapter 14).

Programming Tips

i. R :) :)
ii. Julia :) :)
iii. Python :(
iv. Dynare :), MATLAB :(
v. Stata: Ewww
vi. Maple, Mathematica: Whatever
vii. LaTeX :) :)

Outline of Lecture 1

i. Introduction to Machine Learning
ii. Decision Theory: Loss Functions, Risk, Bayesian approaches: Integrated Risk, Posterior Expected Loss, Bayes' Risk.
iii. Cross Validation and Variance/Bias Trade off
iv. Shrinkage Estimators such as SURE and LASSO
v. Brief Discussion of Economic Application of LASSO if time permits

What is Machine Learning (ML)?

i. Introduction to Machine Learning, Brief History and List of Applications

Relationship Between ML and Econometrics

i. Prediction versus Causation
ii. Mullanaithan paper, Varian Paper

Statistical Decision Theory

i. Decision theory provides a general framework to evaluate desirable properties of various estimators.
ii. It also provides theoretical justification for Bayesian estimators and in what sense they are best or most "reasonable" or "un-dominated".

Examples of Decision Problems

i. Testing the hypothesis that whether graduates with masters degrees earn more in labor market, relative to those without masters or not.
ii. Forecasting recessions by determining recession probabilities, conditional on observable information.
iii. Estimate the causal impact of inflation volatility on income inequality (selfish example, free publicity of my paper :')

Components of General Statistical Decision Problem

i. Observed data $X \in \mathcal{D}$.
ii. Statistical decision function $a = d(X)$, state of world $\omega \in \Omega$.
iii. Loss function $L(a, \omega)$.
iv. Statistical model $f(X|\omega)$ giving conditional probability of data, given the state of world.

Goal of Statistical Decision Theory

i. First, nature draws a state $\omega \in \Omega$, which determines conditional distribution of data X: $f(X|\omega)$.

ii. Decision $a = d(X)$ is taken as function of observable data but *goal is to minimize $L(a, \omega)$*, which depends on what you do and the hidden state.

iii. X is informative about ω through f and hence, informative about L, which is why observing X helps in solving decision problem.

Example of Decision Problem

i. We want to determine probability of recession tomorrow.
ii. ω is joint distribution of relevant business cycle variables over time.
iii. X includes past values of business cycle variables.
iv. $a = d(X)$ is expected probability of recession in next quarter.
v. $L(a, \omega)$ is the expected, economic loss from recession in next quarter in the form of unemployment, GDP loss etc. Why is this expected loss from recession dependent on a?

Defining Loss functions for Estimation

i. We want to do choose $a = d(X)$ to minimize loss $L(a, \omega)$. How to define L?

ii. Common approach is to determine a moment or any other function $\mu(\omega)$, which depends on unobservable state or distribution and minimize distance between a and μ.

iii. For example, $\mu(\omega) = \mathbb{E}[X|\omega] = \int_{X \in \mathcal{D}} X\, f(X|\omega) dX$ is conditional mean which we want to target.

iv. Distance can be squared error loss as also in OLS or absolute deviations, i.e $L(a, \omega) = (a - \mu(\omega))^2$ or $L(a, \omega) = |a - \mu(\omega))|$

Risk Function

i. $R(d,\omega) = \mathbb{E}[L(d(X),\omega)|\omega] = \int_{X\in\mathcal{D}}[L(d(X),\omega)]f(X|\omega)dX$.

ii. Risk is the expected loss, conditional on the true state of nature for any given decision function d.

iii. Estimators or decision functions d which produce lower risk are better.

iv. However, one usually does not have uniform dominance and risk is lower for some ω and higher for other $\bar{\omega} \in \Omega$. Decision theory deals with this key trade off and the variance-bias trade off is an implication of this problem.

Cross-Validation and Variance Bias Trade Off

i. For loss function, $L(a, \omega) = (a - \mu(\omega))^2$

ii. $R(d, \omega) = \mathbb{E}_{\omega}[(d(X) - \mu(\omega))^2] = Var(d(X)) + Bias(d(X))^2$, illustrating variance bias trade off.

Approach 1 for Getting Global Rankings Across Estimators

i. Admissibility and Feasibility, Partial Ordering

Approach 2 for Getting Global Rankings Across Estimators

i. Bayes' Optimality
ii. Integrated Risk, Bayesian Decision Functions
iii. Decision weights and posterior distribution
iv. Posterior Expected Loss and Bayes' Optimal Estimators
v. Complete Ranking

Why Should You be a Bayesian?

i. Complete class theorem
ii. Other? Not subjective probability theory

Regularization/Shrinkage Estimators

i. Regularization is about solving variance bias trade off and this is what shrinkage estimators do (14.2.3 Stachurski).
ii. Shrinkage Estimator 1: Stein's Unbiased Risk Estimate (9.2.3 Stachurski and Kasy Slides).
iii. Shrinkage Estimator 2: LASSO (Tibshirani book and Anders Kock's material may be)

LASSO

i. LASSO

Economic Application of LASSO

i. Big data, dimensionality reduction application from Stock and Watson data set.

Thank you