Statistics 311/Electrical Engineering 377 Project

The final component of the course is to do a course project. The course project should (broadly) be related to some part of the content of the course. The project may be performed in teams of up to three students—though single and pair projects are fine as well.

Please use the course staff—John and Hong—as well as your peers for feedback on the projects. Come talk to us during office hours, and talk early and often. We would love to give you feedback!

1 Types of projects

Here we list a few of the types of projects we expect to see—if you would like to deviate from this, please email the course staff to double check that this is OK. The final product of the research should be a four to six page report detailing your work.

1. *Research project:* original research related to the content of the course. This can be either theoretical work or applied, and ideally will be related to work you are already performing so as to help with your current research interests. These projects could include application of some of the techniques in class to new areas, so that the projects are very applied, or theoretical work.

2. *Pedagogical project:* develop a set of lecture notes and a few problems—of roughly the same difficulty as the problem sets in class—relating to some aspect of the course. A project in this vein may consist of replicating the empirical work in a paper and developing problems associated with it, or surveying an area of statistics and information theory and extracting exercises from a few papers around the area.

2 Logistics and dates

There are three main components for the projects: a proposal, a midterm progress report, and a final report and presentation. You should submit all parts of the project to the course staff at stats311submit@gmail.com. All page limits do *not* include references; we will not count those against the limit.

Part of the project is to give peer review: if you are in the class and doing a project, you will be randomly assigned as a reviewer to two other projects (so that each project will have several reviewers). You will give feedback on these projects throughout the quarter as detailed below.

Proposal: Due Tuesday, 1/26, 5pm. The proposal is a (maximum) two-page document including the names of the student(s) working on the project and describing the planned project. The proposal should include a paragraph of background; a paragraph describing the approach that will be taken; and a paragraph about the hoped for (or desired) results as well as potential failure points of the project.

The proposal is due on Tuesday, 1/26, at 5pm. There are a few components before the proposal is due. By Friday, 1/22, at 6pm, you should email a draft proposal to your peer feedback team—we will provide a list—who will provide feedback on the proposal. In particular, this feedback should address the following criteria:

i. Clarity: is what the project attempting to solve clear? Can anyone who has been in the class for the past few weeks be reasonably expected to understand the material, or at least the introduction?

ii. Relevance: is this project relevant to the course (information theory or statistics)?

iii. Writing style: does the writing have good mathematical notation? Is the grammar acceptable?

You should email your peer feedback by Monday, 1/25, at 8am, cc'ing stats311submit@gmail.com so that we can track the reviews.

Midterm report: Due Tuesday, 2/23 at 5pm. The midterm progress report should be a maximum 4 page document that includes a clear problem statement for your project as well as preliminary results. By the time you write your midterm report, you should have a good idea of your approach.

As with the proposal, peer review is an essential part of the midterm report. By Friday, 2/19 at 6pm, you should email a draft of your midterm report to your feedback team. Peer feedback is due by Monday, 2/22 at 8am, and should cc stats311submit@gmail.com. In addition to feedback on clarity, relevance, and writing style, your feedback should begin with a one paragraph summary of the research.

Submit the midterm progress report, along with all received peer reviews, in a single stapled physical document, by Tuesday 2/23 at 5pm. You should also email the documents to stats311submit@gmail.com.

Final report and presentation: Monday, 3/14, 3pm - 6pm.

Your final report is a maximum six-page document that details your research. The final report should be similar to the midterm progress report, but detailing your approach. Your report should include a clear problem statement for your project, background on the area that can be understood by anyone who has taken the class, and then a description of your approach and results.

We do not require peer review for the final project, though you should feel free to use your peer feedback groups for comments before the final due date.

2.1 Important dates

Friday, 1/22 Draft proposal to peer review teams

Tuesday, 1/26 Submit proposal

Friday, 2/19 Draft midterm report to peer review teams

Tuesday, 2/23 Midterm report

Monday, 3/14 Submit final report, project presentation

Your proposal, midterm report, and final report *must* be written in LaTeX, with appropriate and clear mathematical notation, formulae, etc. This means that mathematics is treated as part of sentences, so that equations end with appropriate punctation.

3 Project ideas

Below we list a few different project ideas; this list is *by no means* exhaustive. Some of these projects may require reading ahead in the lecture notes or reading the lecture notes from last year's offering of the course to be able to complete the project. The course staff can give more information and references on any of the projects if desired. Note that the length of a section does not necessarily correspond to anything we believe to be correlated with the quality of the resulting research; some simply took longer to write.

1. Ranking and collaborative filtering from comparison models. In one version of the collaborative filtering problem (see, for example, the thesis of Oh [14]), we have n users and m items, where the matrix $A \in \mathbb{R}^{n \times m}$ has entries A_{ij} corresponding to user i's rating for item j. When the matrix A is (nearly) low rank, it is possible to accurately reconstruct it from noisy measurements of its entries. In many situations—such as shopping data or web-search—it is easier to recover comparison information (i.e. user i preferred item j to item j') than actual value information, so that it is of interest to recover matrices A under alternate models of observation. Investigate how low-rank recovery types of results may be extended to (noisy) comparison measurements of the matrix A. See also [17, 18, 4, 13, 7].

2. Ranking and collaborative filtering from more advanced comparison models. As in project (1), except that instead of just comparison information, sub-lists of compared items may be observed.

3. Matrix reconstruction with different types of side information. In many ranking or collaborative filtering scenarios, we have access to side information relating objects. For example, in a music recommendation system, we have a matrix $A \in \mathbb{R}^{n \times m}$ where A_{ij} has user i's rating for song j, while a matrix $B \in \mathbb{R}^{m \times d}$ may consist of characteristics of the songs, so that B_{jk} is a measure of how much song j is associated with a genre k. Medical and bio-informatics scenarios also suggest applications of such problems. Investigate matrix recovery under these types of models, either theoretically or by developing algorithmic schemes and applying them.

4. Recent work in theoretical computer science, among other areas, has studied *adaptive inference*, meaning that after performing some computation on a sample or dataset, we ask additional questions. As examples, we might consider high-dimensional inference (see, for example, the survey by Dezeure et al. [8], or papers [23, 12]). In such settings, we have a matrix $X \in \mathbb{R}^{n \times d}$, where $n \ll d$, and observe $Y = X\beta + \varepsilon$, where ε is a mean-zero independent noise variable, and β is sparse, meaning that $\|\beta\|_0 = k \ll d$. We may consider a two-phase procedure in which first we estimate non-zero elements of β, and then in a second (the adaptive phase) phase—after this selection—give confidence intervals or other information on β. As another example, we might collect a sample $X = \{X_1, \ldots, X_n\}$ and evaluate a few hypotheses on the sample, say that functions $\phi_1, \ldots, \phi_m$ are mean-zero. Based on the outputs, we select new functions $\widetilde{\phi}_1, \ldots, \widetilde{\phi}_m$ and wish to test properties of them using the sample. Of course, we have already "used" the sample; there is a natural question of what we can do to avoid over-fitting on the sample in this scenario, which has been studied by Dwork et al. [9, 10] and in an information-theoretic setting by Russo and Zou [21].

   A project in this area could extend the information-theoretic approach of Russo and Zou, give simpler arguments than those by Dwork et al., or provide optimality guarantees or fundamental limits on the number of adaptive questions that can be "asked" of a dataset.

5. In Bayesian statistics, one has a prior belief (represented by a prior distribution π) on a parameter $\theta \in \Theta$, and observes a sample X drawn from a distribution P_θ indexed by the (unknown) θ. Reference analysis (see the survey by Bernardo [3]) advocates choosing a prior π that "allows the data to speak for itself as much as possible," meaning that the prior maximizes the mutual information between the sample X and the parameter θ,

   $$I_\pi(\theta; X) = \int \pi(\theta) p(x \mid \theta) \log \frac{p(x \mid \theta)}{\bar{p}_\pi(x)},$$

   where $\bar{p}_\pi(x) = \int p(x \mid \theta)\pi(\theta)$.

   This choice does not take into account the task at hand, however, including any loss function. Indeed, assume that $L : \Theta \times \Theta \to \mathbb{R}_+$ is a loss function (measuring some performance of an estimate $\widehat{\theta}$ for θ). It is possible to generalize information (as we see later in the class) by looking at the gap between the *prior* risk and *posterior* risk, that is, defining

   $$I_{L,\pi}(\theta; X) := \inf_{\widehat{\theta}} \mathbb{E}_\pi[L(\widehat{\theta}, \theta)] - \mathbb{E}\left[\inf_{\widehat{\theta}} \mathbb{E}[L(\widehat{\theta}, \theta) \mid X]\right] \ge 0.$$

   Investigate the consequences of choosing a prior to maximize this loss-sensitive notion of information. What properties does it have? Are there interesting examples in which it is efficient to compute? Is it robust to prior mis-specification?

6. Priors in online learning (including reinforcement learning and bandit) scenarios. A variety of recent works have studied Thompson-sampling and other Bayesian-based formulations for bandit learning [1]. Many of the convergence guarantees depend on the prior being correctly specified [19, 20]. Investigate the properties of prior mis-specification in such settings.

7. Pedagogical project: explore Good-Turing estimation for different sequences. See, e.g., [15, 16].

8. Learning problems in which the training distribution and testing distribution differ. Can we give optimality results or optimal procedures in this setting? What are good ways to do this? See Ben-David et al. [2].

9. Representation learning and experimental "design." In many machine learning problems, an important object is a function called the kernel, which measures similarity or difference between two objects (i.e. inputs x, x'). A theorem of Bochner is that any kernel that is only a function of the difference $x - x'$ can be written as the characteristic function of a random vector W with (some) distribution P:

   $$k(x, x') = \mathbb{E}_P[e^{iW^\top (x - x')}].$$

   Given a sample, can we efficiently learn a distribution P that generates a kernel with good performance?

10. Distribution property testing and distribution functional estimation: in theoretical computer science and information theory, a recent body of work (see, among many other papers, Jiao et al. [11], Chan et al. [5], and Valiant and Valiant [24]) investigates estimation of quantities such as entropy and testing whether two distributions are close or far apart using a variety of tools. Give a general characterization of these testing (or other) results based on information theoretic quantities such as metric entropy (cf. [25]).

11. Consider the family of probability distributions $\{P_\theta : \theta \in \Theta\}$ parameterized by the decision θ. Let $L : \Theta \times \mathcal{X}$ be a random loss function whose expectation you wish to minimize under P_θ, that is, you wish to minimize $\mathbb{E}_{P_\theta}[L(\theta; X)]$ over θ. This is a setting where your decision θ affects the sampling distribution on X. In many cases, the data comes from a base decision P_0. Then, the corresponding counterfactual risk minimization problem (what would have happened had I sampled from P_θ?) is given as follows:

$$\underset{\theta \in \Theta}{\text{minimize}} \ \mathbb{E}_{P_0}\left[\frac{dP_\theta(X)}{dP_0(X)} L(\theta; X)\right].$$

Often, solving the empirical version of the above optimization problem leads to a overfitted decision in the sense that the resulting estimator $\hat{x}$ has high variance performance. While Swaminathan and Joachims [22] have suggested regularizing by the empirical variance, it has been widely observed in the simulation literature that variance is a poor measure of performance for importance sampling (e.g. Chatterjee and Diaconis [6]). As an alternative, regularizing by other information measures such as $D_{\text{kl}}\left(P_\theta \| P_0\right)$ may be more sensible ([6, Theorem 1.1]). This research project will investigate whether different regularization methods achieve better performance.

References

[1] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Proceedings of the Twenty Fifth Annual Conference on Computational Learning Theory*, 2012.

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[3] J. M. Bernardo. Reference analysis. In D. Day and C. R. Rao, editors, *Bayesian Thinking, Modeling and Computation*, volume 25 of *Handbook of Statistics*, chapter 2, pages 17–90. Elsevier, 2005.

[4] E. J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2008.

[5] S.-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2014.

[6] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. Technical report, Stanford University Department of Statistics, 2015.

[7] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. One-bit matrix completion. *Information and Inference*, page to appear, 2015.

[8] R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: confidence intervals, p-values, and R-software `hdi`. *Statistical Science*, 30(4):533–558, 2015.

[9] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Preserving statistical validity in adaptive data analysis. *arXiv:1411.2664v2 [cs.LG]*, 2014.

[10] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. *arXiv:1506.02629 [cs.LG]*, 2015.

[11] J. Jiao, K. Venkat, Y. Han, and T. Weissman. Minimax estimation of functionals of discrete distributions. *arXiv:1406.6956 [cs.IT]*, 2014.

[12] R. Lockhardt, J. Taylor, R. Tibshirani, and R. Tibshirani. A significance test for the Lasso. *Annals of Statistics*, 42(2):413–468, 2014.

[13] Y. Lu and S. N. Negahban. Individualized rank aggregation using nuclear norm regularization. Technical report, Department of Statistics, Yale University, 2014.

[14] S. Oh. *Matrix Completion: Fundamental Limits and Efficient Algorithms*. PhD thesis, Stanford University, 2010.

[15] A. Orlitsky and A. Suresh. Competitive distribution estimation: Why is Good-Turing good? In *Advances in Neural Information Processing Systems 28*, 2015.

[16] A. Orlitsky, N. Santhanam, and J. Zhang. Always Good-turing: asymptotically optimal probability estimation. In *44th Annual Symposium on Foundations of Computer Science*, 2003.

[17] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[18] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[19] D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *Journal of Machine Learning Research*, page To appear, 2014.

[20] D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems 27*, 2014.

[21] D. Russo and J. Zou. Controlling bias in adaptive data analysis using information theory. *arXiv:1511.05219 [stat.ML]*, 2015.

[22] A. Swaminathan and T. Joachims. Counterfactual risk minimization. In *Proceedings of the 32nd World Wide Web Conference (WWW)*, 2015.

[23] R. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, page To appear, 2014.

[24] G. Valiant and P. Valiant. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems 26*, 2013.

[25] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.