My Research

Predictive and Causal Learning

Niels Richard Hansen

The question that drives my research is: *how do we learn from data*? This is the fundamental question of inductive reasoning, and statistics provides methodology for quantifying how data provides evidence of various positions we may take about "the truth". There is no universally agreed upon correct way of learning from data, but various philosophical positions and a large number of practical models and related methodologies. I seek understanding of the models and the methods we use in all details from a mathematical, a computational and an applied viewpoint.

Learning problems

I will use *precision medicine* to exemplify my current methodological research. The idea of precision or personalized medicine is that treatments of medical conditions can and should be individualized to a much greater extent than is currently the standard. Advocates posit that individualized treatments will have large beneficial effects, and precision medicine is surrounded by a good deal of hype.

To be specific, consider obesity with the body mass index (BMI) as our quantification of how obese an individual is. With data on n adult individuals measuring their BMI together with p *genetic markers* (e.g. SNPs, genetic information represented by a p-dimensional vector $\mathbf{x} = (x_j)$), we can ask questions like: which genetic markers are associated with BMI; can we predict BMI from the genetic markers; and can we design a treatment (a diet, say) based on the genetic markers that will lower BMI for the obese?

The last question on treatment strategies is by far the most difficult question. It is a causal question that probability models cannot answer by themselves. Answering such causal questions is at the core of precision medicine, as we want not only to passively predict the BMI but actively to intervene to affect BMI. My current research is very much focussed on causality, and how we can learn causal structures from partially observed systems.

Prediction of BMI from the composition of genetic markers is a regression problem, and it is a routine statistical problem for $p \sim 10$, but in many ways a non-trivial problem for large p that requires variable selection, incorporation of prior knowledge, and care in the statistical assessments of the learned prediction model.

A causal model is also a predictive model that should remain predictive under various interventions, that is, when certain variables are actively manipulated like changing an individuals diet. A predictive model should "just" generalize beyond the data sample to samples from a reference population. I will restrict attention to predictive learning and explain some recent mathematical result on the assessment of how well a predictive model generalizes.

Predictive model assessment

If y_i denotes BMI for the ith individual in our sample we want a model of the conditional mean, $\mu_i = \mu(\mathbf{x}_i)$, of y_i as a function of the vector $\mathbf{x}_i$ of markers. There is a large number of methods for computing an estimate, $\hat{\mu} \in \mathbb{R}^n$, from the data. Some are linear maps of $\mathbf{y} = (y_i)$ (projections and ridge regression), some are nonlinear but globally Lipschitz (lasso), but many that show good empirical predictive generalization have severe discontinuities (boosted regression trees, random forests, forward stepwise variable selection, relaxed and adaptive lasso etc). The lack of smoothness in $\mathbf{y}$ – or continuity in the first place – has made the mathematical analysis of such estimation methods difficult. Without theoretical results to support practice, there has been an unfortunate tendency in applied statistics of ignoring discontinuities arising from e.g. variable selection, with a resulting overoptimistic assessment of what has been learned from data.

One quantification of predictive performance is the average mean squared prediction error

$$\mathrm{Err} = \frac{1}{n}\sum_i E(Y_i - \hat{\mu}_i)^2$$

for new independent Y_is. It is tempting to use the *training error*

$$\mathrm{err} = \frac{1}{n}\sum_i (y_i - \hat{\mu}_i)^2$$

as an estimate of Err, but this will generally be a downward biased estimate. Thus the training error is overly optimistic about how well the genetic markers predict BMI. Sufficiently imaginative algorithms may easily find models with zero training error even if BMI is, in fact, independent of the markers. In such a case, what we appear to have learned from data is clearly wrong.

The optimism, $\mathrm{Err} - E(\mathrm{err})$, of the training error is rather easy to characterize for linear estimators of μ and also for Lipschitz estimators in a Gaussian model. This can in turn be used to correct err of its bias leading to such statistics as Mallows' C_p (a close relative of AIC) and Stein's unbiased risk estimate (SURE), which are not systematically misleading. A practical deficit of these statistics is that they don't work correctly for discontinuous estimators.

Under the assumption that $\mathbf{Y} \sim \mathcal{N}(\mu, \sigma^2 I)$, Alexander Sokol, Frederik Vissing Mikkelsen and I have in three papers developed representations of $\mathrm{Err} - E(\mathrm{err})$ and computable estimates of this optimism for a number of discontinuous estimators. The representations are all of the form

$$\mathrm{Err} - E(\mathrm{err}) = \frac{2\sigma^2}{n}\left(E(\nabla \cdot \hat{\mu}) + \int \psi \mathrm{d}\nu\right)$$

where ψ is the density for the $\mathcal{N}(\mu, \sigma^2 I)$ distribution and ν is a measure singular w.r.t. Lebesgue measure. Our contribution consists of the term $\int \psi \mathrm{d}\nu$, whose presence is intimately connected to the discontinuities of the map $\mathbf{y} \mapsto \hat{\mu}(\mathbf{y})$.

Alexander Sokol and I considered estimators that are *metric projections* onto closed sets. They are Lipschitz if and only if the set is convex, and thus our results covered such novel examples as q-quasinorm constraints for $q < 1$. The measure ν is always a positive measure, and we were able to obtain bounds, though explicit representations of ν were not obtained, nor did we derive estimates of $\int \psi \mathrm{d}\nu$.

Frederik Vissing Mikkelsen and I used different techniques for estimators that are locally Lipschitz on open subsets $U_i \subseteq \mathbb{R}^n$ with finite perimeter boundaries and such that $\mathbb{R}^n = \cup_i \overline{U}_i$. With $\hat{\mu}^i$ denoting the estimator on U_i we showed that

$$\nu = \frac{1}{2}\sum_{i \neq j} 1_{\overline{U}_i \cap \overline{U}_j} \langle \hat{\mu}^j - \hat{\mu}^i, \eta_i \rangle \cdot \mathcal{H}^{n-1}$$

where η_i denotes the outer unit normal to the boundary of U_i, and $\mathcal{H}^{n-1}$ the $(n-1)$-dimensional Hausdorff measure.

In the latter setting, we used the above representation to show that

$$\int \psi \mathrm{d}\nu^t = \partial_t E(H(t))$$

when $U_i^t = F(t, U_i^0)$ and $\hat{\mu}^{i,t}$ are parametrized by $t \in \mathbb{R}$ and F is a *flow*. Here $t \mapsto H(t)$ is a jump function given in terms of the unit normals, η_i, of U_i^0, the estimators $\hat{\mu}^{i,t}$ and the vector field associated with the flow. Many estimators used in practice can be brought on this form with t a *tuning parameter*. The upshot is that

$$\widehat{\mathrm{Err}}(t) = \mathrm{err}(t) + \frac{2\hat{\sigma}^2}{n}\Big(\nabla \cdot \hat{\mu}^t + \partial_t \mathrm{smooth}(H(t))\Big)$$

is a computable estimate of $\mathrm{Err}(t)$ for the estimator $\hat{\mu}^t$, which can be used to e.g. select the tuning parameter t.