Stochastic optimisation, decision theory and machine learning

- stochastic optimisation and sample average approximation,
- some criteria to evaluate the quality of the estimators,
- decision and statistical learning theories,
- an example - regression.

D. Semenovich, NICTA and University of New South Wales

Stochastic optimisation

$$\min_{x \in \mathcal{X}} \quad \mathbf{E}_\theta \, F(x, \omega) = \int_{\omega \in \Omega} F(x, \omega) p(\omega; \theta) \, d\omega$$

- objective function $F(x, \omega)$ depends on the random variable ω with known distribution $p(\omega; \theta)$, representing uncertainty in measurement, operation or manufacturing processes, computational difficulties.

- denote a solution of the problem as $x^\star_\theta$ as it depends on the distribution $p(\omega; \theta)$.

- constraint set $\mathcal{X}$ can also depend on ω, but I omit this for simplicity.

Sample average approximation

- generate n realisations $(\omega_1, \ldots, \omega_n)$ of "scenarios" or "training data"; we can now form sample average approximations of $\mathbf{E}_\theta F(x, \omega)$:

$$\hat{F}(x, \omega_1, \ldots, \omega_n) = \frac{1}{n} \sum_{j=1}^{n} F(x, \omega_j).$$

- and solve the approximate problem:

$$\hat{x}(\omega_1, \ldots, \omega_n) = \operatorname*{argmin}_{x \in \mathcal{X}} \hat{F}(x, \omega_1, \ldots, \omega_n)$$

How to evaluate this procedure? Note that we at this point have neither the "true solution" nor can we evaluate the "true" objective value $\mathbf{E}_\theta F\big(\hat{x}(\omega_1, \ldots, \omega_n), \omega\big)$.

Decision theoretic optimality criteria for "estimators"

$$R_n(\hat{x}, \theta) = \mathbf{E}_\theta \, F(\hat{x}(\omega_1, \ldots, \omega_n), \omega_{n+1})$$

$$= \int_{\omega^{n+1} \in \Omega^{n+1}} F(\hat{x}(\omega_1, \ldots, \omega_n), \omega_{n+1}) p(\omega_1, \ldots, \omega_{n+1}; \theta) \, d\omega_1 \ldots \omega_{n+1}$$

also known as "frequentist risk", measures "average" performance when trained on "average" data; depends on the distribution $p(\omega; \theta)$. Broadly, two ways to reduce this to a number from a function of θ:

$$R_n^{\text{worst}}(\hat{x}, \Theta) = \max_{\theta \in \Theta} \; R_n \; (\hat{x}, \theta)$$

$$R_n^{\text{Bayes}}(\hat{x}, \pi) = \mathbf{E}_\pi R_n \; (\hat{x}, \theta) = \int_{\theta \in \Theta} R_n(\hat{x}, \theta) \pi(\theta) \, d\theta$$

primarily a tool for analysis (e.g. let $n \to \infty$), but a few closed form or at least computationally tractable solutions (e.g. LQR theory in control, "robust" LPs).

Relation to the classical decision theory problem

Potentially due to historical trajectory, textbook treatments of decision theory focus on the decision variables/parameters. In the stochastic optimisation setting this would look like:

$$R_n(\hat{x}, \theta) = \mathbf{E}_\theta \ell(\hat{x}(\omega_1, \ldots, \omega_n), x^\star_\theta)$$

with the focus on the differences between the estimates $\hat{x}$ and the "true parameters" $x^\star_\theta$.

Statistical learning theory

A way to enrich the basic decision theoretic criteria is to consider "regret" or "excess loss" with respect to the "true" solution over some constraint set $\mathcal{X}$:

$$R_n^{\text{regret}}(\hat{x}, \Theta, \mathcal{X}) = \max_{\theta \in \Theta} \left(R_n\ (\hat{x}, \theta) - \min_{x \in \mathcal{X}}\ \mathbf{E}_\theta\, F(x, \omega) \right)$$

Controlling the size/"capacity"' of set $\mathcal{X}$ (complexity of schedules, smoothness of regression functions etc) allows to devise estimators/decision rules $\hat{x}$ that "work" over larger sets of distributions Θ with respect to this relativised objective.

Example: regression

- $(a, b) \in \mathbf{R}^k \times \mathbf{R}$ have some joint distribution $p\big((a, b); \theta\big)$,
- find weight vector $x \in \mathbf{R}^k$ for which $x^T a$ is a good estimator of b,
- choose x to minimize expected value of the squared loss:

$$\mathbf{E}_\theta \, F(x, (a, b)) = \mathbf{E}_\theta (x^T a - b)^2$$

- we have "training data" or "scenarios" from the joint distribution $(a_i, b_i), \ i = 1, \ldots, n$,
- form an approximate problem using "training data" and denote its

solution by $\hat{x}\big((a_1, b_1), \ldots, (a_n, b_n)\big)$:

$$\hat{x} = \operatorname*{argmin}_x \frac{1}{n} \sum_{i=1}^{n} (x^T a_i - b_i)^2$$

- evaluate sample average approximation of the objective for the model $\hat{x}$ on a new set of m samples $(a'_i, b'_i),\ i = 1, \ldots, m$:

$$\hat{F}\big(\hat{x}, (a'_1, b'_1), \ldots, (a'_m, b'_m)\big) = \frac{1}{m} \sum_{i=1}^{m} (\hat{x}^T a'_i - b'_i)^2$$

  This is essentially "test set" error in machine learning.

Conclusions

- the same setting studied largely independently in different times and communities.
- helpful to attempt to describe same procedures in different vocabularies.
- when can we design "optimal" procedures by mechanically solving optimisation problems?
- the latter would necessitate explicit specification of (often unverifiable) assumptions, rather than vague appeals to LLNs etc.
- any natural ways to control "capacity" of stochastic optimisation problems?