SC4/SM8 Advanced Topics in Statistical Machine Learning

Bayesian Optimisation

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/atsml/
Optimizing “black-box” functions

Machine learning models often have a number of hyperparameters which need to be tuned:

- **kernel methods**: bandwidth in a Gaussian kernel, degree of a Matérn kernel, regularization parameters
- **neural networks**: number of layers, regularization parameters, layer size, batch size, learning rate
- **Latent Dirichlet Allocation**: Dirichlet parameters, number of topics, vocabulary size

Define an objective function: a measure of generalization performance for a given set of hyperparameters obtained e.g. using cross-validation.

- Grid search, random search, trial-and-error...
We are interested in optimizing a 'well behaved' function $f : \mathcal{X} \rightarrow \mathbb{R}$ over some bounded domain $\mathcal{X} \subset \mathbb{R}^d$, i.e. in solving

$$x^* = \arg\min_{x \in \mathcal{X}} f(x).$$

However, $f$ is not known explicitly, i.e. it is a black-box function and we can only ever obtain noisy (and potentially expensive as they may correspond to training of a large machine learning model or even running a complex physical experiment) evaluations of $f$. 
Bayesian Optimisation

**Probabilistic model for the objective $f$**

- Assuming that $f$ is well behaved, we build a surrogate probabilistic model for it (Gaussian Process).
- Compute the posterior predictive distribution of $f$
- Optimize a cheap proxy / acquisition function instead of $f$ which takes into account predicted values of $f$ at new points as well as the uncertainty in those predictions: this model is typically much cheaper to evaluate than the actual objective $f$.
- Evaluate the objective $f$ at the optimum of the proxy.

The proxy / acquisition function should balance exploration against exploitation.
Bayesian Optimisation

Surrogate GP model

Assume that the noise \( \epsilon_i \) in the evaluations of the black-box function is i.i.d. \( \mathcal{N}(0, \delta^2) \):

\[
\begin{align*}
  f & \sim \mathcal{N}(0, \mathbf{K}) \\
  y|f & \sim \mathcal{N}(f, \delta^2 \mathbf{I}).
\end{align*}
\]

Gives a closed form expression for the **posterior predictive mean** \( \mu(x) \) and the **posterior predictive marginal standard deviation** \( \sigma(x) = \sqrt{\kappa(x,x)} \) at any new location \( x \), i.e.

\[
    f(x) | \mathcal{D} \sim \mathcal{N}(\mu(x), \kappa(x,x)),
\]

where

\[
\begin{align*}
  \mu(x) &= \mathbf{k}_{xx}(\mathbf{K} + \delta^2 \mathbf{I})^{-1} \mathbf{y}, \\
  \kappa(x,x) &= k(x,x) - \mathbf{k}_{xx}(\mathbf{K} + \delta^2 \mathbf{I})^{-1} \mathbf{k}_{xx}
\end{align*}
\]

- **Exploitation**: seeking locations with low posterior mean \( \mu(x) \),
- **Exploration**: seeking locations with high posterior variance \( \kappa(x,x) \).
Acquisition functions

- **GP-LCB**. “optimism in the phase of uncertainty”; minimize the lower \((1 - \alpha)\)-credible bound of the posterior of the unknown function values \(f(x)\), i.e.
\[
\alpha_{\text{LCB}} (x) = \mu (x) - z_{1-\alpha} \sigma (x),
\]
where \(z_{1-\alpha} = \Phi^{-1} (1 - \alpha)\) is the desired quantile of the standard normal distribution.

- **PI** (probability of improvement). \(\tilde{x}\): the optimal location so far, \(\tilde{y}\): the observed minimum. Let \(u (x) = 1 \{ f (x) < \tilde{y} \}\),
\[
\alpha_{\text{PI}} (x) = \mathbb{E} [u(x) | \mathcal{D}] = \Phi (\gamma (x)) , \quad \gamma (x) = \frac{\tilde{y} - \mu (x)}{\sigma (x)}
\]

- **EI** (expected improvement). Let \(u (x) = \max (0, \tilde{y} - f (x))\)
\[
\alpha_{\text{EI}} (x) = \mathbb{E} [u(x) | \mathcal{D}] = \sigma (x) (\gamma (x) \Phi (\gamma (x)) + \phi (\gamma (x))) .
\]

Treating \(\tilde{y}\) as the actual value \(f(\tilde{x})\) of the objective?
Illustrating Bayesian Optimization

slides from A Tutorial on Bayesian Optimization for Machine Learning by Ryan Adams
Illustrating Bayesian Optimization

slides from A Tutorial on Bayesian Optimization for Machine Learning by Ryan Adams
Illustrating Bayesian Optimization

slides from *A Tutorial on Bayesian Optimization for Machine Learning* by Ryan Adams
Illustrating Bayesian Optimization

slides from A Tutorial on Bayesian Optimization for Machine Learning by Ryan Adams
Illustrating Bayesian Optimization

slides from "A Tutorial on Bayesian Optimization for Machine Learning" by Ryan Adams
Illustrating Bayesian Optimization

slides from "A Tutorial on Bayesian Optimization for Machine Learning" by Ryan Adams
slides from A Tutorial on Bayesian Optimization for Machine Learning by Ryan Adams
We considered a selection of topics in statistical machine learning, but there is much more!

- **Topics we did not cover**: multitask learning, online learning, reinforcement learning, deep learning, message passing algorithms (belief propagation, expectation propagation), variational autoencoders, generative adversarial networks, ensemble methods, boosting, causality, interpretability, fairness,...

- **Further resources**:
  - Bishop, Pattern Recognition and Machine Learning, Springer.
  - Murphy, Machine Learning: A Probabilistic Perspective, MIT Press.
  - Shalev-Shwartz and Ben-David, Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press.
  - Schölkopf and Smola, Learning with Kernels, MIT Press.
  - Rasmussen and Williams, Gaussian Processes for Machine Learning, MIT Press.
  - Goodfellow, Bengio and Courville, Deep Learning, MIT Press.
  - Machine Learning Summer Schools, videolectures.net.
  - Conferences: NIPS, ICML, AISTATS, UAI.

- Please fill in the [online feedback form](#).