Machine Learning in Three Lectures

Sach Mukherjee

DZNE, Bonn

Lecture III: Some connections, frontier topics and the philosophy of ML

- Unifying themes: some connections between ML methodologies that may historically have been studied independently
- Selected frontier topics: a (highly selective) view of a few areas of current research
- The mindset of ML and its social and scientific outlook

Some connections

(1) Bayesian inference and regularization

- We saw several examples of constrained optimization used to e.g. allow modelling in high dimensions
- Close connection between such optimization-based methods and Bayesian inference
- First, a quick introduction to Bayesian inference...

Bayesian inference (in one slide)

- Assume data X from probability model f with unknown parameters θ
- *Likelihood* is joint probability of data given parameter, i.e. $f(X \mid \theta)$
- View parameter itself as RV, encode "pre-data" knowledge via *prior distribution* $\pi(\theta)$
- *Posterior distribution* describes knowledge about θ after seeing the data
$$p(\theta \mid X) = \frac{1}{Z} f(X \mid \theta)\pi(\theta)$$
- This allows uncertainty quantification etc. (but may be difficult to access in practice for high-dimensional θ)
- *Maximum a posterori* or MAP estimate is posterior mode, i.e.
$$\hat{\theta}_{\mathrm{MAP}} = \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid X)$$

Linear regression revisited

- Recall linear regression model: vector of n outputs Y, p-dimensional inputs X and p-dimensional parameter β
- Probability model: $Y_i \mid X_i, \beta, \sigma \sim N(\beta^{\mathrm{T}} X_i, \sigma^2)$
- Likelihood $p(Y_1 \ldots Y_n \mid X, \beta, \sigma^2) = \prod_i N(Y_i \mid \beta^{\mathrm{T}} X_i, \sigma^2)$
- Set $\pi(\beta) = N(0, c)$ to discourage extreme values
- Posterior

$$p(\beta, \sigma \mid X, Y) \propto p(Y \mid X, \beta, \sigma)\, \pi(\beta \mid \sigma)\, \pi(\sigma)$$

Regularized estimation and Bayesian models

- Posterior proportional to $p(Y \mid X, \beta, \sigma)\,\pi(\beta \mid \sigma)\,\pi(\sigma)$
- Taking logs and maximizing wrt β

$$
\begin{aligned}
\hat{\beta}_{\mathrm{MAP}} &= \operatorname*{argmax}_{\theta} p(\beta, \sigma \mid X, Y) \\
&= \operatorname*{argmin}_{\theta} \underbrace{\|Y - X\beta\|_2^2}_{\textit{log likelihood}} + \underbrace{\lambda\|\beta\|_2^2}_{\textit{log prior}}
\end{aligned}
$$

- Identical to ridge regression! Quadratic regularizer can be viewed as coming from the Gaussian prior on β

- More generally, close connection to formal probability models, constrained optimization can be viewed as MAP under some prior
- This is both illuminating and useful in designing and understanding penalties

(2) PCA and K-means as matrix factorizations

- PCA can also be viewed as an approximate factorization of the data matrix of the form $X \approx ZU$ and interpreted as solving the least squares problem
$$J(Z, U) = \|X - ZU\|_F^2$$
s.t. identifiability constraint on U
- Recall that K-means models the data as K groups each with its own mean μ_k. Collect μ_k's together into a $K \times p$ matrix U. The K-means objective sums distances to corresponding means and can be written as
$$J(Z, U) = \|X - ZU\|_F^2$$
s.t. constraint on Z (binary with exactly one non-zero in each row)
- Thus, PCA and K-means are both matrix factorizations of the data, subject to different constraints
- Many other matrix factorizations are possible, aimed at different goals, leading to different objectives

(3) Decision theory and optimal model complexity

- Consider again the classification problem with two classes
- Suppose the groups are relatively well separated but that the class-specific distributions are complicated (e.g. full covariance Gaussians, or some more complicated high-dimensional density).
- Let the true model be $f(\cdot;\theta^*)$ and some simpler model (e.g. the linear one) be $g(\cdot;\psi)$. Assume g is wrong in the sense that $\nexists\psi : g(\cdot;\psi) = f(\cdot;\theta^*)$
- (For a given task/loss function) should we use the correct model class f or the incorrect one g?

All models are wrong, some are useful

- The expected loss for the fitted model depends on the initial dataset $D_n = \{X_i, Y_i\}_{i=1\ldots n}$ via the parameter estimates
- If the initial data is fixed, and parameters are estimated from these data, we can write the expected loss under the correct model as

$$R_n(\hat{f}) = \mathbb{E}[L(Y, f(X; \hat{\theta}(D_n)))]$$

and under the incorrect one as

$$R_n(\hat{g}) = \mathbb{E}[L(Y, g(X; \hat{\psi}(D_n)))]$$

- In general, there is no guarantee that $R_n(\hat{f}) < R_n(\hat{g})$! This depends on an interplay between the true data-generating process, the model classes f and g and the properties of the estimators $\hat{\theta}$ and $\hat{\psi}$
- Easy to sketch examples for binary classification

Unifying ML models

- There are many more connections between ML methods that may have been developed independently for seemingly different tasks
- Many models can be viewed: heuristically via an intuitive objective; as motivated by a probability model; from a Bayesian viewpoint; and so on
- Regularization is a key theme for high dimensional data and the art is in defining suitable schemes that impose appropriate structure for specific problems

Selected frontier topics

(1) Networks and causality

- Often data X come from systems where the connections between the variables are of interest and not just predicting one output or finding groups
- Network models are a natural tool. A class of statistical models called *graphical models* are very widely used
- These allow potentially high-dimensional joint distributions (over components of X) to be studied from the point of view of conditional independence structure. Examples include Markov random fields and Bayesian networks
- General idea is to factor the (potentially) large joint distribution using a graph G, which is itself often of scientific interest

Networks and causality

- In Gaussian case zeros in the inverse covariance are non-edges in the graph (these give conditional independence relationships)
- There are now many schemes for fitting high-dimensional models (not necessarily Gaussian) using regularization as seen before
- Example: the *graphical lasso* estimates $\Omega = \Sigma^{-1}$ as

$$\hat{\Omega} = \underset{\Omega}{\text{argmax}} \; \underbrace{\log|\Omega| - \mathrm{Tr}(n^{-1}X^{\mathrm{T}}X\Omega)}_{\textit{Likelihood}} - \underbrace{\lambda\|\Omega\|_1}_{\textit{Penalty}}$$

  This induces sparsity directly in the inverse and is a scalable and powerful way to model in high dimensions
- However, graphical models are models of joint statistical distributions and do not necessarily have a *casual* interpretation

Networks and causality

- A long-term goal remains to extract causal insights from data
- Much progress, we now know this is possible, at least to some extent, in some circumstances
- General problem is *confounding* – many correlations may be due to shared causation by some other factor (the confounder)
- Current efforts focus on a range of ways to extract causal insights from high-dimensional data and to test such procedures in a systematic way

(2) Deep learning

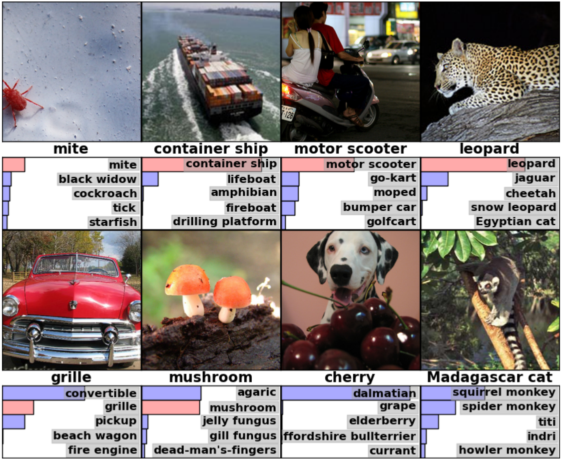
- A neural network (NN) is a composition of layers each of which is a linear function pushed through a nonlinearity and which overall maps an input X to an output Y
- Thus, a NN learns a function $Y = f(X;\theta)$ where the overall f is nonlinear and θ collects all model parameters
- Deep neural networks (DNNs) have many layers (hence "deep"). These are very flexible models with typically ultra-high-dimensional θ
- In applications like computer vision intermediate layers can be seen to learn "features" at different levels

Deep learning

- Recall, ML method = (model, estimator, computation). All three aspects have been critical for making DNNs work
- There have been several breakthroughs in fitting DNNs based on a combination of innovations in the models, computing and regularization coupled with many heuristics and engineering details
- DNNs have made huge gains in a number of specific tasks, notably image classification and also in some non-obvious settings where the models are used provide a data representation that then allows for greater generalization

Deep learning – some history

- Historically, after much interest in early days of ML, much of the community essentially stopped working on NNs, and most applications – from image analysis to speech recognition – settled on other approaches
- But performance on standardized "challenges" made the community sit up and take notice
- On an image classification challenge, error rates dropped *dramatically* in 2012 (and are now even lower). This was entirely unprecedented
- Achieved using DNNs with $\sim 6 \times 10^7$ parameters (and a lot of regularization). Illustrated again the *statistical* power of flexible models with sufficient regularization



(Krizhevsky et al., NIPS, 2012)

Consider also again the "captioning" example, and think about how large the output set is and how difficult this is *de novo*...

man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

(Karpathy & Li, CVPR, 2015)

DNNs reinforce again importance of model complexity, representation learning and empirical risk

(3) Very high dimensional data

- Advances in high-dimensional inference have cumulatively completely changed our view – now clear that it is very much possible to effectively and robustly fit very high-dimensional models, using various forms of regularization
- Hard to overstate how mindset has changed in last ∼15 years or so
- Typically regularization imposes some generic structure – like sparsity – leaving the models to work out the details of *which* sparsity pattern works
- Very different from strong scientific prior knowledge, because the models are general and can be run on problems where little or nothing is understood in a real scientific sense

Very high dimensional data

- High dimensionality can be viewed as a *modelling choice*: is it better to include more (and regularize more strongly) or include less (but run the risk of missing the signal)?
- Sparse methods negotiate the tradeoffs of high-dimensional data automatically

(4) Real world prediction

- As the basic machinery has become more powerful, a range of issues that arise in real problems have become more important
- A model that works well on a given dataset may not generalize to data from a seemingly similar problem due to subtle differences in the underlying distributions
- This *dataset shift* is a bit different to the train/test issue because the new dataset is not even a sample from the same distribution
- Flexible models can end up subtly "tuned" to details that do not transfer to the new setting
- Real-world implications of decisions need to be understood and used to inform/constrain learning

Real world prediction

- *Transfer learning* refers to the task of adapting a model to a new setting without throwing away what was already learned
- *Multi-task learning* is a general term for learning jointly over somehow related outputs/tasks
- Many issues arise in *designing risk estimators*. Questions relating to precisely how training and testing should be done, how models should be monitored and updated and so on

ML: the philosophy, mindset and outlook

The machine learning mindset

- ML can be viewed as a natural outcome of advances in computing coupled with an empirical mindset that gained ground during the 20th century
- The main difference from other fields – even much of statistics or econometrics – is the view of models
- In ML – and much of modern statistics – the idea is to work on problems where there is no satisfactory model (due to scale, complexity etc.)
- Related to fact that understanding causality and predicting specific aspects are obviously related but not identical.

The machine learning mindset

- The idea is that even if the system is very complicated, given suitable data (not necessarily huge), one can detect useful patterns that can essentially substitute for what one would have done with a model
- Almost any current ML application bears this view out – the effectiveness does *not* rest on having a good model in the usual sense
- Thus, the philosophy is that flexible formulations (high-dimensional data, relatively complex models or both) coupled with systematic and empirically-guided regularization can often bring very complex problems – diagnosis, decision making, AI – into reach

ML: the outlook

- It is now clear that ML can transform many processes – this is only in its infancy – and in some cases allow entirely new tasks to be taken on
- The *economics of ML* are remarkable. A fitted model may have low-to-zero marginal cost
- Adaptation to ML has in some ways been slow, much to be done
- Outlook: ML will become much more embedded in society, the economy and science. Important to foster understanding and to have a much broader conversation going forward

Selected textbooks

- MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. CUP
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer
- Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press
- Goodfellow, I., Bengio, Y. and Courville, A. (2017). *Deep Learning*. MIT Press
- Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer