Paper 3a Review: Random Features for Large Scale Kernel Machines

Ali Rahimi, Benjamin Recht

Vegen Soopramanien, vs385

September 26 2019

---

1 Summary

The paper proposes a novel idea: while the kernel trick projects our data into a higher dimensional space such that a linearly separable hyperplane can be learned, the scalability of the Gram matrix with respect to the training size is problematic: instead, Rahimi et al. propose to factor the kernel function itself: they map the data into a low-dimensional random feature space. The authors propose two randomized feature maps to achieve this. The first is a map consisting of sinusoids drawn randomly from the Fourier transform of the kernel function that we seek to approximate. The second map they propose ("binning") does a partitioning of the input space using randomly shifted grids at randomly chosen resolutions. These randomized features have inner products that uniformly approximate many popular kernel functions, including the Gaussian or the Lapacian.

2 Strengths

- The mathematical derivations in both the random Fourier features map and the binning are very clear and easy to follow, involving simple algebra and mathematical analysis. The result of equation (2) is not too hard to derive for instance for the reader.
- The structure is clearly presented: Rahimi et al. explain the issues with the kernel trick, and show how their approach scales well with the training data.
- The algorithms they present are well formulated and so a reader can easily implement those in some framework without much trouble.

3 Weaknesses

- The authors could have spent more time on explaining the core results from their experiments comparing how their novel approaches perform relative to state-of-the-art methods present in the literature (like exact SVM).
- Maybe the authors could also have ran experiments like using their random feature mapping on unsupervised learning methods (maybe on PCA in the same way as kernelized PCA exists)- they did experiments only on classification and regression settings.
- I wouldn't necessarily treat this as a weakness but more of a recommendation to the authors: in a sense, random Fourier features is like initializing a neural network with the cosine function representing the non-linearity: this is a very powerfully intuitive way of thinking about this novel approach that would have been super helpful to any reader had the authors just mentioned it perhaps?

4 Paper's Impact

- The paper opened up a whole new field of technique that rivalizes with kernelized machines' scalability problems.
- The authors left the readers with the possibility to experiment on their own and come up with hybrid versions of the random Fourier features and binning features by concatenating both.
- The computational speedup by first computing random features and then applying the associated linear technique is a win for the industry!