Assorted Topics in Probabilistic ML (2)

CS772A: Probabilistic Machine Learning
Piyush Rai
Plan for today

- Assorted Topics
  - Conformal Prediction (simple and fast way to get prediction uncertainty/set)
  - Nonparametric Bayesian methods (learning the right model size/complexity)
Conformal Prediction

- A simple technique to easily obtain confidence intervals
  - In classification, such an interval may refer to the set of highly likely classes for a test input

- For more difficult test inputs, the set would typically be larger

- In a way, conformal prediction gives predictive uncertainty
  - However, unlike Bayesian ML, we don’t get model uncertainty
  - Only one model is learned in the standard way and we construct the set of likely classes
  - It’s like a black-box method; no change to training procedure for the model
Conformal Prediction

- Assume we already have a trained model \( \hat{f} \) using some labelled data
- Idea: Use a calibration set of \( n \) examples to generate a prediction set \( \mathcal{C}(X_{test}) \) s.t.

\[
1 - \alpha \leq p(Y_{test} \in \mathcal{C}(X_{test})) \leq 1 - \alpha + \frac{1}{n + 1}
\]

The approach* to construct the prediction set \( \mathcal{C}(X_{test}) \) is as follows:
- Assuming classification task, for each example in the calibration set, compute

\[
s_i = 1 - \hat{f}(x_i)_{y_i}
\]
- Compute the \( 1 - \alpha \) quantile of \( s_1, s_2, \ldots, s_n \). Call it \( \hat{q} \)
- Now the calibration set for a new test input \( X_{test} \) can be defined as

\[
\mathcal{C}(X_{test}) = \{y: \hat{f}(X_{test})_y \geq 1 - \hat{q}\}
\]

*A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (Angelopoulos and Bates, 2022)
Nonparametric Bayesian Methods

- Need for nonparametric Bayesian modeling
- Some motivating problems
- NPBayes modeling mixture models (clustering)
- Some standard ways of constructing NPBayes models
  - Stick-breaking process, Dirichlet process
  - Some metaphors: Chinese Restaurant Process
Motivating Problem: Mixture Models

- Suppose each observation is generated from a $K$ component mixture model

\[ z_n \sim \text{multinoulli}(\pi) \]
\[ x_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n}) \]

- How to learn $K$, i.e., the number of components (clusters) for such a mixture model?

- Can use marginal-likelihood based model selection but is expensive
  - Need to train the model several times for each possible value of $K$

- Also difficult if the data is streaming (hard to know beforehand how many clusters)

- How about a prior over $Z = [z_1, z_2, \ldots, z_N]$ (or $\pi$) that allows learning the “right” $K$?
Motivating Problem: Latent Feature Models

- Suppose each observation is a subset sum of \( K \) “basis vectors” (or “latent features”*)

\[
z_{nk} \sim \text{Bernoulli}(\pi_k) \quad k = 1, \ldots, K
\]

\[
x_n = \sum_{k=1}^{K} z_{nk} a_k + \epsilon_n = A z_n + \epsilon_n
\]

- This can also be seen as a special type of matrix factorization \( \mathbf{X} = \mathbf{Z A}^\top + \mathbf{E} \)

- How about a prior over \( \mathbf{Z} \) (or \( \mathbf{A} \) or \( \bm{\pi} = [\pi_1, \ldots, \pi_K] \)) that allows learning the “right” \( K \)?

* Indian Buffet Process: An Introduction and Review (Griffiths and Ghahramani, 2011)
Motivating Problem: SVD-style Matrix Factorization

Consider the following SVD-style decomposition for an $N \times M$ matrix $X$:

$$X = \sum_{k=1}^{K} \lambda_k u_k v_k^T + E = U \Lambda V^T + E$$

- Each $u_k \in \mathbb{R}^N$, $v_k \in \mathbb{R}^M$, $\lambda_k \in \mathbb{R}$, and $\Lambda$ is a $K \times K$ diag matrix with $\lambda_k$'s on diags.

- This is basically a weighted sum of $K$ rank-1 matrices:
  - $\lambda_k$'s are the weights
  - $\lambda_k$'s are akin to the singular values in SVD

- How to learn $K$, i.e., the “rank” of the above factorization?

- How about a prior on $\Lambda$, or $U$ or $V$, that allows us to learn the “right” $K$?

*Sparse Bayesian infinite factor models (Bhattacharya and Dunson, 2011)*
Nonparametric Bayesian Modeling

- Enables constructing models that do not have an *a priori* fixed size

- Nonparametric does not mean no parameters
  - Instead, have a possibly infinite (unbounded) number of parameters
  - Note: We’ve already seen Gaussian Processes which is a nonparametric Bayesian model

- Usually constructed via one of the following ways
  - Take a finite model (e.g., a finite mixture model) and consider its “infinite limit”
  - Have a model that allows very large number of params but has a “shrinkage” effect, e.g.,
    \[ X = \sum_{k=1}^{K} \lambda_k u_k v_k^T + E \]
    \[ \lambda_k \rightarrow 0 \quad \text{as} \quad k \rightarrow \infty \]
  - We will look at some examples of both these approaches

A vast area of research in ML and statistics. We will only be looking at a basic flavor of some approaches

And can potentially grow as we see more and more data (actual number will depend on the amount/properties of data)

A tutorial on Bayesian nonparametric models (Gershman and Blei, 2012)
Being Nonparametric by Taking Infinite Limit of Finite Models
A Finite Mixture Model

- Data \( \mathbf{X} = [x_1, x_2, \ldots, x_N] \), cluster assignments \( \mathbf{Z} = [z_1, z_2, \ldots, z_N] \), \( K \) clusters
- Denote the mixing proportion by a vector \( \boldsymbol{\pi} = [\pi_1, \ldots, \pi_K] \), \( \sum_{k=1}^{K} \pi_k = 1 \)

\[
\begin{align*}
    p(\boldsymbol{\pi} | \alpha) &= \text{Dirichlet} \left( \frac{\alpha}{K}, \frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \right) \\
    p(z_n | \pi) &= \prod_{k=1}^{K} \pi_k^{z_{nk}} \\
    p(\mathbf{X} | \pi) &= \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k p(x_n | z_n = k)
\end{align*}
\]

a.k.a. “collapsing” a variable; one less variable to infer now

- Integrating out \( \boldsymbol{\pi} \), the marginal prior probability of cluster assignments

\[
p(\mathbf{Z} | \alpha) = \int p(\mathbf{Z} | \pi)p(\pi | \alpha) d\pi = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(m_k + \frac{\alpha}{K})}{\Gamma \left( \frac{\alpha}{K} \right)^{K}}
\]

(verify)
A Finite Mixture Model

- The prior distribution of $z_n$ given cluster assignment $Z_{-n}$ of other points?

  $p(z_n|Z_{-n}, \alpha) = \frac{p(z_n, Z_{-n}|\alpha)}{p(Z_{-n}|\alpha)} = \frac{p(Z|\alpha)}{p(Z_{-n}|\alpha)}$

  A discrete distribution (multinoulli) since $z_n$ can take one of $K$ possibilities

- Using $p(Z|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(m_k+\alpha)}{\Gamma(\frac{\alpha}{\alpha})^K}$ we have

  $p(z_n = j|Z_{-n}, \alpha) = \frac{p(z_n = j, Z_{-n}|\alpha)}{p(Z_{-n}|\alpha)} = \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \frac{\Gamma(m_j+\alpha)}{\Gamma(N-1+\alpha)} \prod_{k\neq j} \frac{\Gamma(m_k+\alpha)}{\Gamma(\frac{\alpha}{\alpha})^K}$

  $= \frac{m_{-n}J + \frac{\alpha}{\alpha}}{N - 1 + \alpha}$

  Number of points in cluster $j$, not counting $x_n$

- Note: Can also get this result using $p(z_n = j|Z_{-n}, \alpha) = \int p(z_n = j|\pi)p(\pi|Z_{-n}, \alpha)d\pi$

- Thus prior prob. of $z_n = j$ is proportional to how many other points are in cluster $j$

- Note that it also implies that mixture models have a rich-gets-richer property
  - Meaning: a priori, a cluster with more points is likely to attract more points
Taking the Infinite Limit..

- Since \( p(z_n = j|Z_{-n}, \alpha) = \frac{m_{-n,j} + \frac{\alpha}{K}}{N-1+\alpha} \), as \( K \to \infty \), \( p(z_n = j|Z_{-n}, \alpha) = \frac{m_{-n,j}}{N-1+\alpha} \).

- Suppose only \( K_+ \) clusters are currently occupied (i.e., have at least one data point).

- Total prob. of \( x_n \) going to any of these \( K_+ \) clusters \( = \sum_{j=1}^{K_+} \frac{m_{-n,j}}{N-1+\alpha} = \frac{N-1}{N-1+\alpha} \).

- Probability of \( x_n \) going to a new (i.e., so far unoccupied) cluster \( = \frac{\alpha}{N-1+\alpha} \).

- Therefore in the limit of an unbounded number of clusters, we have

\[
p(z_n = j|Z_{-n}, \alpha) = \begin{cases} \frac{m_{-n,j}}{N-1+\alpha} & \text{(prob. of going to } j = 1, \ldots, K_+ \text{)} \\ \frac{\alpha}{N-1+\alpha} & \text{(prob. of creating a new cluster } K_+ + 1 \text{)} \end{cases}
\]

- The above gives us a prior distribution for mixture models with unbounded \( K \).
  - Can combine it now with the suitable likelihood to infer the posterior* of \( Z \).

- Note: Prob. of starting a new cluster is prop. to Dirichlet hyperparam \( \alpha \) (can learn it).

A Metaphor: Chinese Restaurant Process (CRP)

- Assume a restaurant with infinite number of tables (each table denotes a cluster)
- Customer 1 sits at a randomly chosen table (all tables are equivalent to begin with)
- Each subsequent customer $n > 1$ sits using the following scheme
  - Sits at an already occupied table $k$ with probability $\frac{m_k}{n-1+\alpha}$
  - Sits at a new table with probability $\frac{\alpha}{n-1+\alpha}$
Being Nonparametric using Models that have a Shrinkage Effect
Consider a finite mixture model with $K$ components with params $(\mu_k, \Sigma_k)_{k=1}^K$.

In the finite case, we can assume $\pi = [\pi_1, \ldots, \pi_K]$ and $\pi \sim \text{Dirichlet} \left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$.

We can make it a nonparametric model by making $\pi$ an infinite-dimensional vector.

How to construct such a vector? Is there an infinite dimensional Dirichlet distribution?
Mixture Models: Two Equivalent Views

Prior (a.k.a. “base distribution” for the parameters of each mixture component)

Example: $G_0$ can be NIW if each component is a Gaussian and $\phi_k = (\mu_k, \Sigma_k)$

No explicit cluster ids; instead, $\theta_i$ denotes the param of the distribution which will generate $x_i$

Since $G$ is discrete, there will at most be $K$ distinct $\theta_i$’s, thereby achieving clustering

Typical way of showing the plate notation of a mixture model

$\pi \sim \text{Dirichlet} \left( \frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \right)$

$\phi_k \sim G_0$  \hspace{1cm} $k = 1, 2, \ldots, K$

$z_i \sim \text{multinoulli}(\pi)$  \hspace{1cm} $i = 1, 2, \ldots, N$

$x_i \sim p(x | \phi_{z_i})$  \hspace{1cm} $i = 1, 2, \ldots, N$

$\pi \sim \text{Dirichlet} \left( \frac{\alpha}{K}, \ldots, \frac{\alpha}{K} \right)$

$\phi_k \sim G_0$  \hspace{1cm} $k = 1, 2, \ldots, K$

$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$

$\theta_i \sim G$  \hspace{1cm} $i = 1, 2, \ldots, N$

$x_i \sim p(x | \theta_i)$  \hspace{1cm} $i = 1, 2, \ldots, N$

But how to construct such a $G$ distribution with potentially infinite components?
Stick-Breaking Process (Sethuraman’94)

- Recursively break a length 1 stick into two pieces
- Assume breaking point in each round is drawn from a Beta distribution

\[ \beta_k \sim \text{Beta}(1, \alpha) \quad k = 1, \ldots, \infty \]

\[ \pi_1 = \frac{\beta_1}{k-1} \]

\[ \pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell) \quad k = 2, \ldots, \infty \]

- Can show that \( \sum_{k=1}^{\infty} \pi_k - 1 \to 0 \) which is what we want

- We can now have a “nonparametric/infinite” mixture distribution

\[ G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \]

“Location/atoms” \( \phi_k \) can be drawn from a “base” distr \( G_0 \), say NIW if

\[ \phi_k = (\mu_k, \Sigma_k) \]

- We basically replaced the Dirichlet prior on \( \pi \) by a Stick-Breaking Process (SBP) prior

SBP gives us a way to construct infinite dimensional Dirichlet distribution known as the “Dirichlet Process”
Infinite Dimensional Dirichlet

- Drawing from an infinite-dim Dirichlet would give an infinite-dim prob. vector
  \[ \mathbf{\pi} = [\pi_1, \pi_2, \pi_3, \ldots] \]

- We can construct this vector to have very few entries as nonzero

- Consider recursively drawing from a Dirichlet as defined below

\[
\begin{align*}
1 & \sim \text{Dirichlet}(\alpha) \\
(\pi_1, \pi_2) & \sim \text{Dirichlet}(\alpha/2, \alpha/2) \\
(\pi_1\pi_{11}, \pi_1\pi_{12}, \pi_2\pi_{21}, \pi_2\pi_{22}) & \sim \text{Dirichlet}(\alpha/4, \alpha/4, \alpha/4, \alpha/4)
\end{align*}
\]

As the concentration parameter gets smaller and smaller, the split of values in LHS get more and more skewed.

Therefore, after doing the above a few times, the \( \mathbf{\pi} \) vector will only have a very few entries as nonzero and in the infinite-sized \( \mathbf{\pi} \), there will only be a finite many nonzero entries, and rest will be zero.

This is basically what happens in the case of Dirichlet Process / Stick-Breaking Process.
Dirichlet Process - Formally

- A Dirichlet Process $DP(\alpha, G_0)$ defines a distribution over distributions
  - So $G \sim DP(\alpha, G_0)$ will give us a distribution
  - $\alpha$: concentration param, $G_0$: base distribution (\text{mean of DP})
  - Large $\alpha$ means $G \rightarrow G_0$

- **Fact 1:** If $G \sim DP(\alpha, G_0)$ then any finite dim. marginal of $G$ is Dirichlet distributed
  \[
  [G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))
  \]
  for any finite partition $A_1, \ldots, A_K$ of the space $\Omega$ (Ferguson, 1973)

- **Fact 2:** Any $G$ drawn from $DP(\alpha, G_0)$ will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)

- **Fact 3:** $G$ is a discrete dist, i.e., only a few $\pi_k$'s will be significant
Summary

- We saw an example of a nonparametric Bayesian model
  - CRP/Dirichlet Process: For clustering problems
- NPBayes models exist for many other problems, e.g., matrix factorization
- Many applications of these models to solve a wide range of problems
- Also saw GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (both supervised and unsup. learning)
- Rich theory based on stochastic processes (beyond the scope of this course)
- Inspired other non-probabilistic algos, e.g., Using Dirichlet Process Mixture Model to get a $K$-means like clustering algorithm (DP-means) which doesn’t require $K$