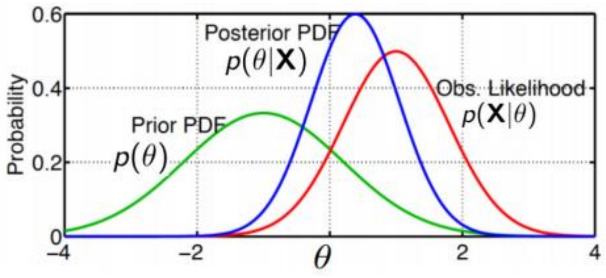
Bayesian Inference for Some Simple Models

CS698X: Topics in Probabilistic Modeling and Inference
Piyush Rai

Recap: Bayesian Inference

- Given data $\mathbf{X}$ from a model $\boldsymbol{m}$ with parameters $\boldsymbol{\theta}$, the posterior over θ

$$p(\theta|\mathbf{X},m) = \frac{p(\mathbf{X}|\theta,m)p(\theta|m)}{p(\mathbf{X}|m)} = \frac{p(\mathbf{X}|\theta,m)p(\theta|m)}{\int p(\mathbf{X}|\theta,m)p(\theta|m)d\theta} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$



- Can use the posterior for various purposes, e.g.,
  - Getting point estimates e.g., mode (though direct point estimation is easier)
  - Uncertainty in our estimates of $\boldsymbol{\theta}$ (variance, credible intervals, etc)
  - Computing the posterior predictive distribution (PPD) for new data, e.g.,

$$p(\boldsymbol{x}_*|\mathbf{X},m) = \int p(\boldsymbol{x}_*|\theta,m)p(\theta|\mathbf{X},m)\, d\theta$$

Often a useful way to compute PPD for some models without finding the posterior explicitly

Another interesting interpretation: PPD is the ratio of two marginal likelihoods

$$p(\boldsymbol{x}_*|\mathbf{X},m) = \frac{p(\boldsymbol{x}_*,\mathbf{X}|m)}{p(\boldsymbol{X}|m)}$$

Marginalization using the posterior distribution of $\boldsymbol{\theta}$

Equivalent to marginalizing $\boldsymbol{\theta}$ from the plug-in predictive

- Caveat: Computing posterior/PPD is in general hard (due to the intractable integrals)

Recap: Marginal Likelihood and Its Usefulness

- Likelihood vs Marginal Likelihood: $p(\mathbf{X}|\theta, m)$ vs $p(\mathbf{X}|m)$
  - Prob. of $\mathbf{X}$ for a single θ under model m vs prob. of $\mathbf{X}$ averaged over all θ's under model m
- Can use marg. lik. $p(\mathbf{X}|m)$ to select the best model from a finite set of models

$$\hat{m} = \arg\max_m p(m|\mathbf{X}) = \arg\max_m p(\mathbf{X}|m)p(m) = \arg\max_m p(\mathbf{X}|m), \text{ if } p(m) \text{ is uniform}$$

- Also useful for estimating hyperparam of a model (if m denotes hyperparams)
  - Suppose hyperparams of likelihood are α_ℓ and that of prior are α_p (so here $m = \{\alpha_\ell, \alpha_p\}$)
  - Assuming prior $p(\alpha_\ell, \alpha_p)$ is uniform, hyperparams can be estimated via MLE-II

$$\{\hat{\alpha}_\ell, \hat{\alpha}_p\} = \arg\max_{\alpha_\ell, \alpha_p} p(\mathbf{X}|\alpha_\ell, \alpha_p) = \arg\max_{\alpha_\ell, \alpha_p} \int p(\mathbf{X}|\theta, \alpha_\ell)p(\theta|\alpha_p)d\theta$$

  - Again, note that the integral here may be intractable and may need to be approximated
- Can also compute model posterior $p(m|\mathbf{X})$ and do Bayesian Model Averaging

$$p(\boldsymbol{x}_*|\mathbf{X}) = \sum_{m=1}^{M} p(\boldsymbol{x}_*|\mathbf{X}, m)p(m|\mathbf{X})$$

Bayesian Inference for Multinoulli/Multinomial

- Assume N discrete obs $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$ with each $x_n \in \{1,2,\dots,K\}$, e.g.,
  - x_n represents the outcome of a dice roll with K faces
  - x_n represents the class label of the n^{th} example in a classification problem (total K classes)
  - x_n represents the identity of the n^{th} word in a sequence of words

These sum to 1

- Assume likelihood to be multinoulli with unknown params $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_K]$

$$p(x_n|\pi) = \text{multinoulli}(x_n|\pi) = \prod_{k=1}^{K} \pi_k^{\mathbb{I}[x_n=k]}$$

Generalization of Bernoulli to $K > 2$ discrete outcomes

- $\boldsymbol{\pi}$ is a vector of probabilities ("probability vector"), e.g.,
  - Biases of the K sides of the dice
  - Prior class probabilities in multi-class classification $(p(y_n = k) = \pi_k)$
  - Probabilities of observing each word of the K words in a vocabulary

Called the concentration parameter of the Dirichlet (assumed known for now)

Large values of α will give a Dirichlet peaked around its mean (next slide illustrates this)

Each $\alpha_k \geq 0$

- Assume a conjugate prior (Dirichlet) on $\boldsymbol{\pi}$ with hyperparams $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_K]$

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1, \dots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1}$$

Generalization of Beta to K-dimensional probability vectors

Brief Detour: Dirichlet Distribution

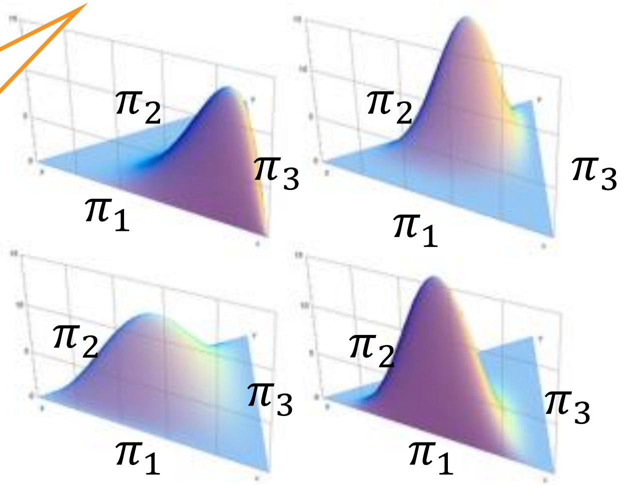
Basically, probability vectors

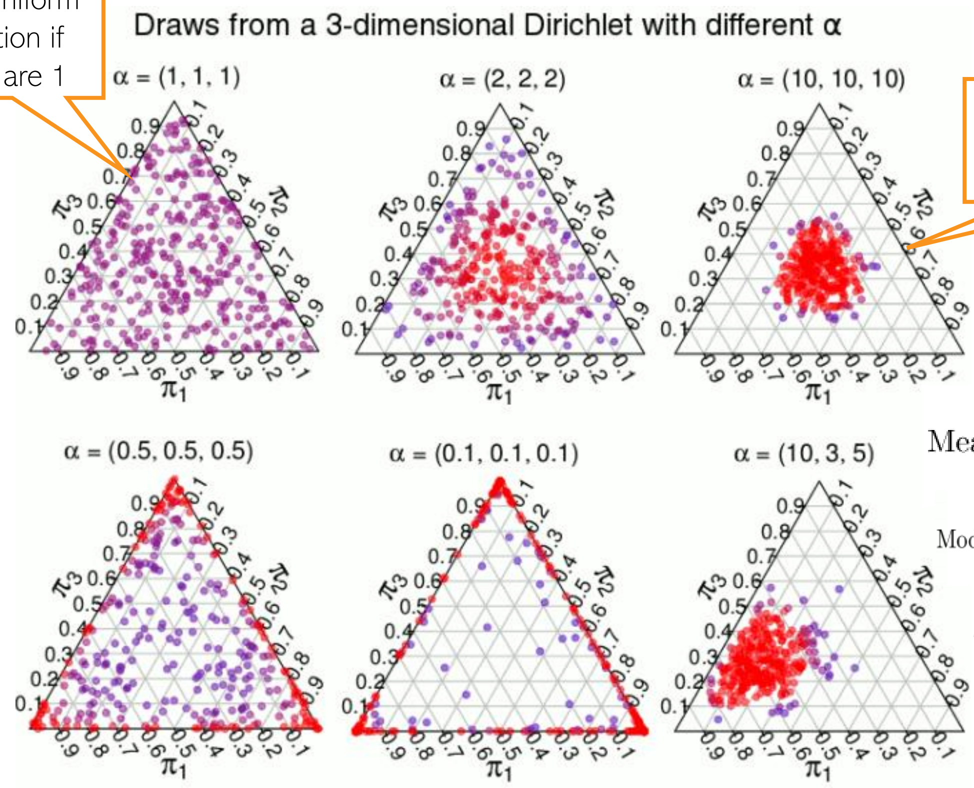
- An important distribution. Models non-neg. vectors $\boldsymbol{\pi}$ that also sum to one
- A random draw from K-dim Dirich. will be a point under $(K$-1$)$-dim probability simplex

Like a uniform distribution if all α_k's are 1

Visualizations of PDFs of some 3-dim Dirichlet distributions (each generated using a different conc. Param vector $\boldsymbol{\alpha}$)

$\boldsymbol{\alpha}$ controls the shape of the Dirichlet (just like Beta distribution's hyperparameters)





All α_k's large results in peak around the center of the simplex

$$\text{Mean} = \left[\frac{\alpha_1}{\sum_{k=1}^K \alpha_k}, \dots, \frac{\alpha_K}{\sum_{k=1}^K \alpha_k}\right]$$

$$\text{Mode} = \left[\frac{\alpha_1 - 1}{\sum_{k=1}^K \alpha_k - K}, \dots, \frac{\alpha_K - 1}{\sum_{k=1}^K \alpha_k - K}\right] \quad (\alpha_k > 1)$$

$$\text{var}(\pi_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \quad \alpha_0 = \sum_{k=1}^K \alpha_k$$

- Interesting fact: Can generate a K-dim Dirichlet random variable by independently generating K gamma random variables and normalizing them to sum to 1

Bayesian Inference for Multinoulli

Likelihood

Prior

- Posterior $p(\boldsymbol{\pi}|\mathbf{X})$ is easy to compute due to conjugacy b/w multinoulli and Dir.

$$p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha}) = \frac{p(\boldsymbol{\pi},\mathbf{X}|\boldsymbol{\alpha})}{p(\mathbf{X}|\boldsymbol{\alpha})} = \frac{p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\alpha})}{p(\mathbf{X}|\boldsymbol{\alpha})} = \frac{p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\mathbf{X}|\boldsymbol{\pi})}{p(\mathbf{X}|\boldsymbol{\alpha})}$$

Don't need to compute for this case because of conjugacy

Marg-lik = $\int p(\boldsymbol{\pi}|\boldsymbol{\alpha})p(\mathbf{X}|\boldsymbol{\pi})d\boldsymbol{\pi}$

- Assuming x_n's are i.i.d. given $\boldsymbol{\pi}$, $p(\mathbf{X}|\boldsymbol{\pi}) = \prod_{n=1}^N p(x_n|\boldsymbol{\pi})$, and therefore

$$p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha}) \propto \prod_{k=1}^K \pi_k^{\alpha_k - 1} \times \prod_{n=1}^N \prod_{k=1}^K \pi_k^{\mathbb{I}[x_n=k]} = \prod_{k=1}^K \pi_k^{\alpha_k + \sum_{n=1}^N \mathbb{I}[x_n=k] - 1}$$

- Even without computing marg-lik, $p(\mathbf{X}|\boldsymbol{\alpha})$, we can see that the posterior is Dirichlet
- Denoting $N_k = \sum_{n=1}^N \mathbb{I}[x_n = k]$, number of observations with with value k

$$p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$$

Similar to number of heads and tails for the coin bias estimation problem

- Note: $N_1, , N_2 \ldots, N_K$ are the sufficient statistics for this estimation problem
  - We only need the suff-stats to estimate the parameters and values of individual observations aren't needed (another property from exponential family of distributions – more on this later)

Bayesian Inference for Multinoulli

- Finally, let's also look at the posterior predictive distribution for this model
- PPD is the prob distr of a new $x_* \in \{1,2,\dots,K\}$, given training data $\mathbf{X} = \{x_1, x_2, \dots, x_N\}$

Will be a multinoulli. Just need to estimate the probabilities of each of the K outcomes

$$p(\boldsymbol{x}_*|\mathbf{X},\boldsymbol{\alpha}) = \int p(\boldsymbol{x}_*|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha})d\boldsymbol{\pi}$$

- $p(\boldsymbol{x}_*|\boldsymbol{\pi}) = \text{multinoulli}(\boldsymbol{x}_*|\boldsymbol{\pi})$, $p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha}) = \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1+N_1,\alpha_2+N_2,\dots,\alpha_K+N_K)$
- Can compute the posterior predictive probability for each of the K possible outcomes

$$p(\boldsymbol{x}_* = k|\mathbf{X},\boldsymbol{\alpha}) = \int p(\boldsymbol{x}_* = k|\boldsymbol{\pi})p(\boldsymbol{\pi}|\mathbf{X},\boldsymbol{\alpha})d\boldsymbol{\pi}$$

$$= \int \pi_k \times \text{Dirichlet}(\boldsymbol{\pi}|\alpha_1+N_1,\alpha_2+N_2,\dots,\alpha_K+N_K)d\pi$$

$$= \frac{\alpha_k + N_k}{\sum_{k=1}^K \alpha_k + N}$$ (Expectation of π_k w.r.t the Dirichlet posterior)

- Thus PPD is multinoulli with probability vector $\left\{\frac{\alpha_k+N_k}{\sum_{k=1}^K \alpha_k+N}\right\}_{k=1}^K$

Note how these probabilities have been "smoothened" due to the use of the prior + the averaging over the posterior

A similar effect was achieved in the Beta-Bernoulli model, too

- Plug-in predictive will also be multinoulli but with prob vector given by the point estimate of $\boldsymbol{\pi}$

Applications?

- Beta-Bernoulli and Dirichlet-Multinoulli/Multinomial models are widely used
- Now know how to do fully Bayesian inference (or point estimation) if our model has such sub-components, and how to compute plug-in/posterior predictive distributions

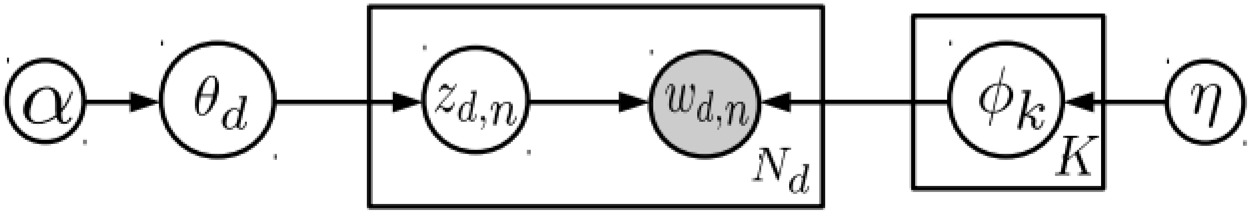


Plate-diagram of the Latent Dirichlet Allocation (LDA) topic model (will see it later)

- Some popular examples are
  - Models for text data: Each document can be modeled as a bag-of-words (Beta-Bernoulli) or a sequence of token (Dirichlet-Multinoulli)
  - Bayesian inference for class prior probabilities in generative classification models: Class labels of training examples are observations and class prior probabilities are to be estimated
  - Bayesian inference for mixture models: Cluster ids are our (latent) “observations” of Dir-Mult model and mixing proportions are to be estimated
  - .. and several others