Basics of Parameter Estimation in Probabilistic Models

CS698X: Topics in Probabilistic Modeling and Inference

Piyush Rai

Two Fundamental Rules

- Keep in mind these two simple rules of probability: sum rule and product rule
- Assume two random variables $\boldsymbol{a}$ and $\boldsymbol{b}$

$$p(a) = \sum_b p(a,b) \qquad \textbf{(sum rule)}$$

$$p(a,b) = p(a)p(b|a) = p(b)p(a|b) \qquad \textbf{(product rule)}$$

- Note: For continuous r.v.'s, sum replaced by integral: $\boldsymbol{p(a) = \int p(a,b)db}$
- Bayes rule can be easily obtained from the above two rules
- Assuming $\boldsymbol{b}$ is continuous, the Bayes rule is

$$p(b|a) = \frac{p(b)p(a|b)}{p(a)} = \frac{p(b)p(a|b)}{\int p(a,b)db} = \frac{p(b)p(a|b)}{\int p(b)p(a|b)db}$$

- Probabilistic modeling and inference is about consistently applying these two rules

Probabilistic Modeling

- Assume data $\mathbf{X} = \{\boldsymbol{x}_n\}_{n=1}^{N}$ generated from a prob distribution with params $\boldsymbol{\theta}$

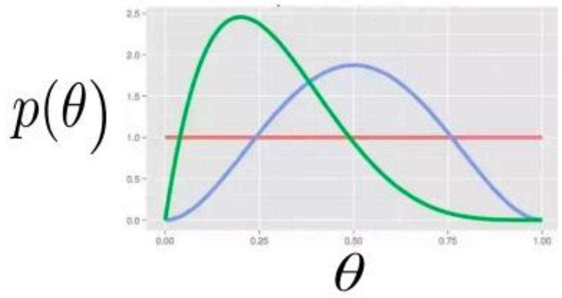
$$\boldsymbol{x}_n \sim p(\boldsymbol{x}|\theta, m) \qquad\qquad n = 1,2, \dots, N$$

- $p(\boldsymbol{x}|\theta, m)$ is also known as the likelihood (a function of the parameters $\boldsymbol{\theta}$)
- Assume a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{m})$ on the parameters $\boldsymbol{\theta}$
- Note: Here $\boldsymbol{m}$ collectively denotes "all other stuff" about the model, e.g.,
  - An "index" for the type of model being considered (e.g., "Gaussian", "Student-t", etc)
  - Any other (hyper)parameters of the likelihood/prior
- Note: Usually we will omit the explicit use of $\boldsymbol{m}$ in the notation
  - In some situations (e.g., when doing model comparison/selection), we will use it explicitly
- Note: For some models, the likelihood is not defined explicitly using a probability distribution but implicitly† via a probabilistic simulation process

†Hierarchical Implicit Models and Likelihood-Free Variational Inference (Tran et al (NIPS 2017))

Probabilistic Modeling

- The prior $p(\theta|m)$ plays an important role in probabilistic/Bayesian modeling
  - Reflects our prior beliefs about possible parameter values before seeing the data

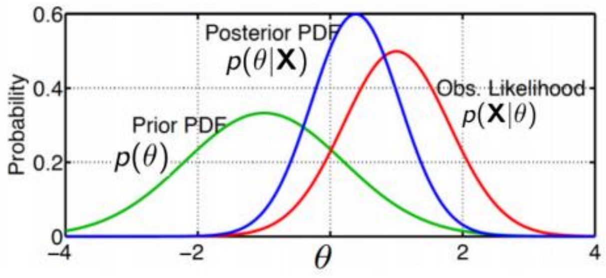


  - Can be "subjective" or "objective" (also a topic of debate, which we won't get into)
  - Subjective: Prior (our beliefs) derived from past experiments
  - Objective: Prior represents "neutral knowledge" (e.g.. uniform, vague prior)
  - Can also be seen as a regularizer (connection with non-probabilistic view)
- The goal of probabilistic modeling is usually one or more of the following
  - Infer the unknowns/parameters $\boldsymbol{\theta}$ given data $\mathbf{X}$ (to summarize/understand the data)
  - Use the inferred quantities to make predictions

Parameter Estimation/Inference

- Can infer params by computing posterior distribution (fully Bayesian inference)

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



Note: Prior and posterior are distributions over θ. Likelihood is just a function of θ

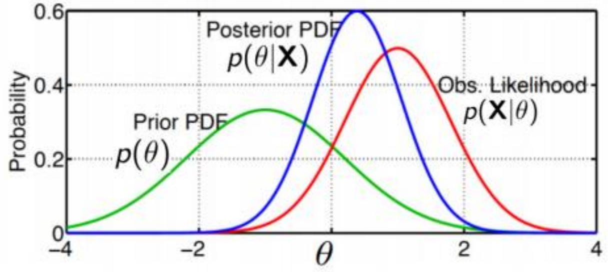
- Marginal likelihood is another very important quantity (more on it later)
  - Probability of data after integrating out some/all of the unknowns from the likelihood
  - $p(\mathbf{X}|m)$ above is the likelihood obtained after integrating out θ from the likelihood $p(\mathbf{X}|\theta,m)$
  - Not always available in closed form (the key reason why full posterior is often hard to compute)
- Cheaper alternative to fully Bayesian inference: Point Estimation of the parameters
  - Find the single "best" estimate of the unknowns

Point Estimation

In some problems as we will see, hybrid inference is also possible/desirable – infer full posterior for some parameters and point estimate for others

- Recall that the posterior is

$$p(\theta|\mathbf{X}, m) = \frac{p(\mathbf{X}|\theta, m)p(\theta|m)}{p(\mathbf{X}|m)}$$



- Point estimation typically done via one of the following two approaches
  - **Maximum likelihood (ML) estimation:** Find θ for which <u>observed data has largest probability</u>

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \ \log p(\mathbf{X}|\theta)$$

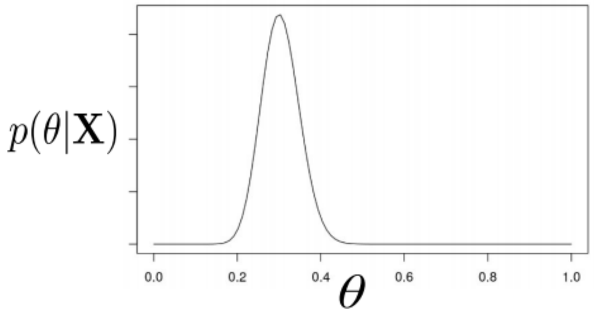
  - **Maximum a posteriori (MAP) estimation:** Find θ that has the <u>largest posterior probability</u>

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \ \log p(\theta|\mathbf{X}) = \underset{\theta}{\operatorname{argmax}} \left[\log p(\mathbf{X}|\theta) + \log p(\theta)\right]$$

- MAP is just like MLE but information from the prior is also incorporated
  - Thus MAP is like regularized MLE (thus helps prevent overfitting)

"Reading" the Posterior Distribution

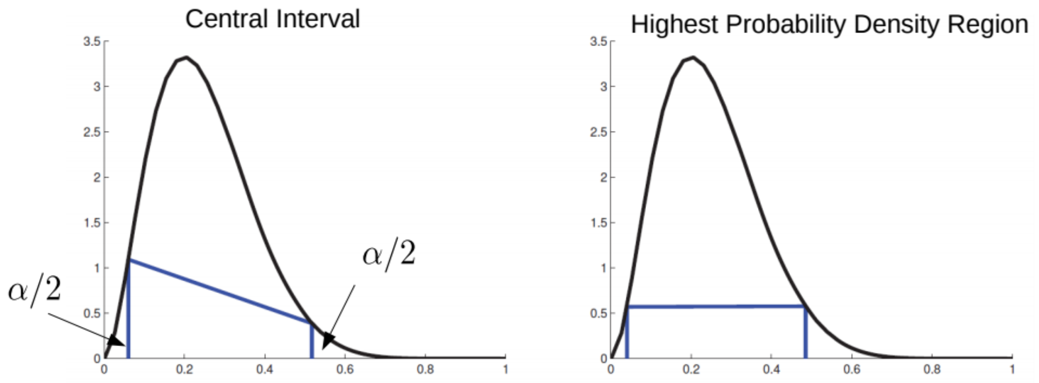
- Posterior provides us a holistic view about $\boldsymbol{\theta}$ given observed data
- A simple unimodal posterior for a scalar parameter $\boldsymbol{\theta}$ might look something like

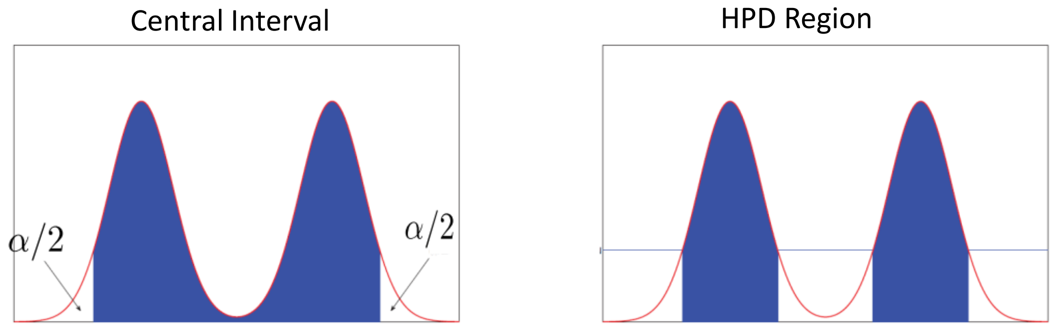


- Various types of estimates regarding $\boldsymbol{\theta}$ can be obtained from the posterior, e.g.,
  - Mode of the posterior (same as the MAP estimate)
  - Mean and median of the posterior
  - Variance/spread of the posterior (uncertainty in our estimate of the parameters)
  - Any quantile (say $\mathbf{0} < \alpha < \mathbf{1}$ quantile) of the posterior, e.g., $\boldsymbol{\theta}_*$ s.t. $\boldsymbol{p}(\boldsymbol{\theta} \leq \boldsymbol{\theta}_*) = \boldsymbol{\alpha}$
  - Various types of intervals/regions

“Reading” the Posterior Distribution

Also defined for multi-modal posteriors





- $100(1-\alpha)\%$ Credible Interval: Region in which $1-\alpha$ fraction of posterior's mass resides

$$\mathcal{C}_\alpha(\mathbf{X}) = (\ell, u) : p(\ell \le \theta \le u|\mathbf{X}) = 1-\alpha$$

Computing central interval or HPD usually requires inverting CDFs

- Credible Interval is not unique (there can be many $100(1-\alpha)\%$ intervals)
- Central Interval is a symmetrized version of Credible Interval ($\alpha/2$ mass on each tail)
- Another useful interval: The $(1-\alpha)$ Highest Probability Density (HPD) region

$$\mathcal{C}_\alpha(\mathbf{X}) = \{\theta : p(\theta|\mathbf{X}) \ge p^*\} \quad \text{s.t.} \quad 1-\alpha = \int_{\theta:p(\theta|\mathbf{X})\ge p^*} p(\theta|\mathbf{X})d\theta$$

Using Posterior for Making Predictions

- Posterior can be used to compute the posterior predictive distribution (PPD)
- PPD is essentially our test time prediction using the learned model
- The PPD of a new observation $\boldsymbol{x}_*$ given previous observations

Prediction by averaging over the posterior distribution of the unknowns parameters

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

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

Just a simple example. The actual form of PPD (e.g., what we are predicting and what we condition on, etc) will depend on the problem.

Assuming observations are i.i.d. given θ

This integral is only rarely tractable

- Computing PPD requires doing a posterior-weighted averaging over all values of $\boldsymbol{\theta}$
- A crude approximation: Instead of PPD, just use a plug-in predictive

However, this ignores all the uncertainty about $\boldsymbol{\theta}$

$$p(\boldsymbol{x}_*|\mathbf{X},m) \approx p\left(\boldsymbol{x}_*\middle|\hat{\theta},m\right)$$

Here $\hat{\theta}$ is the ML or MAP estimate of the parameters

- Plug-in pred. is the same as PPD with $p(\theta|\mathbf{X},m)$ approximated by a point mass at $\hat{\theta}$
  - If we are using plug-in predictive, we are not really being Bayesian!

Marginal Likelihood

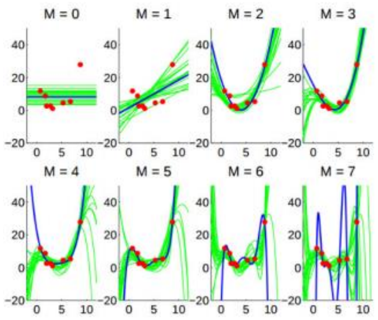
- Recall the Bayes rule for computing the posterior

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

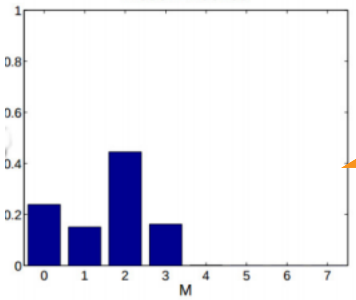
- The denominator in the Bayes rule is the marginal likelihood (a.k.a. "model evidence")
- Marginal lik. is the same as expected likelihood (exp. under the prior distribution) since

$$p(\mathbf{X}|m) = \int p(\mathbf{X}|\theta,m)p(\theta|m)d\theta = \mathbb{E}_{p(\theta|m)}[p(\mathbf{X}|\theta,m)]$$

- For a good model $\boldsymbol{m}$, we would expect marg. lik. to be large (most $\boldsymbol{\theta}$'s will be good)
  - Can thus compare two models $\boldsymbol{m}$ and $\boldsymbol{m'}$ by comparing the respective marg. lik.



$p(\mathbf{X}|m)$



This doesn't require a separate validation set unlike cross-validation

Model Selection and Model Averaging

- Marginal likelihood is hard-to-compute (due to integral) but a very useful quantity
- It can be used for doing model selection
  - Choose model $m \in \{1,2,\dots,M\}$ that has largest posterior probability

$$\hat{m} = \arg\max_m p(m|\mathbf{X}) = \arg\max_m \frac{p(\mathbf{X}|m)p(m)}{p(\mathbf{X})} = \arg\max_m p(\mathbf{X}|m)p(m)$$

Then, for prediction, we can report the PPD $p(\boldsymbol{x}_*|\mathbf{X},\hat{m})$ of the best model $\hat{m}$

That is, simply comparing the marginal likelihoods

  - Note: If all models are equally likely a priori then $\hat{m} = \arg\max_m p(\mathbf{X}|m)$
  - Note: If m denotes a hyperparam, then $\hat{m} = \arg\max_m p(\mathbf{X}|m)$ is the optimal hyperparameter
    - Called MLE-II for hyperparameter estimation (find hyperparams that maximize the marginal prob. of data)
- Using the model posterior $p(m|\mathbf{X})$, we can even average over models

Posterior based averaging over all models $m = 1,2,\dots,M$ and all possible param of each model

Called Bayesian Model Averaging (BMA)

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

PPD of $\boldsymbol{x}_*$ under model m

Posterior probability of model m

- Since $p(\boldsymbol{x}_*|\mathbf{X},m) = \int p(\boldsymbol{x}_*|\theta,m)p(\theta|\mathbf{X},m)d\theta$ BMA is like double averaging to make prediction

Coming Up Next

- Some simple examples of parameter estimation in probabilistic models
  - Estimating the bias of a coin given previous outcomes of tosses from a Bernoulli model
  - Estimating the mean of a Gaussian given observations from a Gaussian model