Today: ML in graph model
Bayesian methods
Model selection

Estimation in GM

DGM: parametric family $P_{\theta} = \{ p(x) = \frac{1}{Z(\theta)} p(x|\theta) \}$
$\theta = (\theta_1, \ldots, \theta_M)$
$\theta = \theta_1 \times \cdots \times \theta_M$  \hspace{1cm} M = p

IE: no tying of parameters

$\Rightarrow$ MLE decouples in $p$ independent problems

$\frac{1}{n} \sum_{i=1}^{n} p(\text{data} | \theta) = \frac{1}{n} \prod_{i=1}^{n} p(x^{(i)} | \theta) = \prod_{i=1}^{p} \frac{1}{\sum_{j=1}^{p} p(x^{(i)} | x^{(i)}_j, \theta)}$

$\log \left[ \frac{1}{p} \sum_{j=1}^{p} \left( \frac{1}{\sum_{i=1}^{n} \log p(x^{(i)}_j | x^{(i)}_j, \theta) - \sum_{i=1}^{n} \log p(x^{(i)} | x^{(i)}_j, \theta) \right) \right]$

Example: for discrete H.V. $\Rightarrow \theta^{(n)} = \frac{\#(x_i = k, x_j = \text{something})}{\#(x_j = \text{something})}$

\textbf{If} latent variable (undesired)

$\Rightarrow$ use EM

Undi. GM: example for exponential family:

$p(x | \theta) = \exp \left( \sum_{i=1}^{C} n_i^C T_i(\theta) - \Lambda(\theta) \right)$

$\exp(\sum_{i=1}^{C} n_i^C T_i(\theta)) = \Psi_C(\theta)$

Gradient ascent on log-likelihood

$\frac{1}{n} \sum_{i=1}^{n} \log p(x^{(i)} | \theta) = \sum_{i=1}^{n} \log \left( \sum_{i=1}^{C} \frac{1}{n} n_i^C T_i(x^{(i)}) \right) - \frac{1}{n} \Lambda(\theta)$

$\nabla_{\theta}(\theta) = \hat{\mu} - \frac{\sum_{i=1}^{n} T_i(\theta)}{p(x | \theta)}$
Bayesian methods:

Data $X \sim \theta$ and Model $\theta$ gives $X \sim \text{Bernoulli}(\theta)$

"Frequentist" vs. "Bayesian"

Bayesian: $p(\theta|X) \propto p(X|\theta)p(\theta)$

Frequentist: $p(X|\theta)$

Subjective Bayesian $\rightarrow$ use probability everywhere that is uncertain

Frequentist $\rightarrow$ use analysis tools

Example: biased coin

Bayesian model $\theta \sim \text{Beta}(\alpha, \beta)$

Posterior $\theta|X \sim \text{Beta}(\alpha + \sum x_i, \beta + n - \sum x_i)$

Questions: What is prob. of next flip = 1?
Bayesian: integrates out uncertainty

\[ p(x_{n+1} | x_1:n) = \int p(x_{n+1} | \theta) p(\theta | x_1:n) \, d\theta \]

\[ p(x_{n+1} | x_1:n) = \int \theta p(\theta | x_1:n) \, d\theta \rightarrow \text{posterior mean?} \]

\[ E[\theta | \text{data}] = \frac{\alpha}{\alpha + \beta} = \frac{n + 1}{n + 2} \]

\[ \hat{\theta}_{\text{posterior mean}} = \frac{\alpha}{\alpha + \beta} \left[ \frac{n}{n + 2} \right] + \frac{1}{2} \left[ \frac{2}{n + 2} \right] \left[ 1 - \alpha \mu_n \right] \]

\[ \mu_n \xrightarrow{\text{in}} \mu^* \]

\[ \text{Variance of a Beta} \ (\alpha + \beta)^2 (\alpha + \beta + 1) \]

\[ = \frac{(n + 1)}{n + 2} \left( 1 - \frac{n}{n + 2} \right) \frac{2}{n + 2} \xrightarrow{n \to \infty} 0 \]

posterior "contracts" around \( \hat{\theta}_n \xrightarrow{\text{as}} \theta^* \) true value

\[ \hat{\theta} \xrightarrow{\text{as}} \theta^* \]

"Bernstein von-Mises theorem"

\[ \rightarrow \text{"Bayesian CLT"} \]

basically says that if prior placed

\[ \text{large mass around true model } \theta^* \]

from posterior concentrates around \( \hat{\theta} \) as

a Gaussian asymptotically

recall from week 1: multinomial model

\[ x \sim \text{Mult}(\theta, k) \quad \text{where } \theta \in \Delta_k \]

\[ \hat{\theta}_{\text{ML}} = \frac{x_k}{n} \]

Putting Dirichlet prior over \( \theta \),

\[ \theta \sim \text{Dir}(\alpha) \]

we also get Dirichlet posterior \( \theta | \text{data} \sim \text{Dir}(\alpha + x_k + k) \)

\[ \rightarrow \text{Dirichlet-Gibbs is a conjugate prior for Multinomial likelihood} \]
Consider family of dist: \( F = \{ p(\theta | \alpha) : \alpha \in \mathbb{R}^2 \} \)

Say \( F \) is "conjugate family" to observation model \( p(x | \theta) \)

If posterior \( p(\theta | x, \alpha) \in F \)

\[ \exists \alpha' s.t. \quad p(\theta | x, \alpha) = p(\theta | x, \alpha') \]

\[ p(x | \theta) \frac{p(\theta | \alpha)}{\int p(\theta | \alpha') d\alpha} \]

Side note: If use conjugate family pairs in a DBM, then Gibbs sampling can be easy.

**Model selection:**

\[ M_1 \quad M_2 \]

\( \text{(note here } M_1 \subseteq M_2 \text{)} \)

As a frequentist:

\[ \hat{\theta}^M_{M_1} = \arg \max \log p(\text{data} | \theta_1, \theta_2, \text{model} = M_1) \]

\[ \hat{\theta}^M_{M_2} = \arg \max \log p(\text{data} | \theta_1, \theta_2, \text{model} = M_2) \]

Different space

Is \( \log p(\text{data} | \hat{\theta}^M_{M_1}, M = M_1) \) vs. \( \log p(\text{data} | \hat{\theta}^M_{M_2}, M = M_2) \)?

Since \( M_1 \subseteq M_2 \rightarrow \text{LHS} \leq \text{RHS} \)

i.e. cannot use likelihood directly to choose model

\(- \text{instead, use cross-validation} \text{ (i.e. } \log p(\text{valid} | \hat{\theta}^M_{M_1(\text{train}), M = M_1}) \text{)} \)

**Bayesian alternative:**

True Bayesian \( \rightarrow \) sum over models (integrate out uncertainty)

Introduce prior over models \( p(M) \)

\[ p(\text{new} | \mathcal{D}) = \sum_M p(\text{new} | M, \mathcal{D}) p(M | \mathcal{D}) d\theta \]
\[
\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(D|M_1)p(M_1)}{p(D|M_2)p(M_2)} \quad \text{Bayes factor}
\]

Choosing between \( k \) models \( M_1, \ldots, M_k \):

Pick \( M_i \) which maximizes \( p(\text{data}|M=M_i) \),

"empirical Bayes"  
"type II ML"

Is this a good thing? It is better than ML because \( p(\text{data}|M=M_0) \) is normalized over the possible datasets.

\[
\left[ \text{vs. } \frac{p(\text{data}|S_{\text{ml}}(\text{data}), M=M_i)}{p(\text{data}|S_{\text{ml}}(\text{data}), M=M_j)} \text{ which could be high on all datasets} \right]
\]

can too much. Suppose \( M_1 \leq M_2 \leq M_3 \) (increasing complexity): \( p(D|M=M_3) = \text{uniform on some support} \)
Type II ML can still overfit if have many models e.g. say $P(D|M) = S(D|M)$

```
\[ \text{data} \]
```

how to compute marginal likelihood ?

usually need approximations \[ \text{via} \text{Markov chain Monte Carlo} \]

**Bayesian information criterion**

is a (rough) approximation of 

\[
\log P(\text{data} | M) \approx \log P(\text{data} | \hat{\theta}_M) - \frac{d \log(n)}{2} \\
\text{complexity penalty}
\]

use Laplace approximation

\[
P(D | M) = \int \frac{d\theta}{\exp(-n h(\theta))} P(D | \theta) P(\theta | M) \, d\theta
\]

where 

\[
h(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i | \theta, M) + \log P(\theta | M)
\]

\[ \text{do Taylor expansion of this around } \hat{\theta}_M \]

2 approximations : keep only term which grows with $n$. get BIC

BIC is "consistent"