Bayesian Learning, Factor Models and Approximate Inference

02459 Machine Learning for Signal Processing

Ole Winther
IMM, DTU B321, Off. 115
owi@imm.dtu.dk

Outline

- Student exercise: Where are the cancers?
- Learning from data I: Setting-up and learning the model.
- Sly gambler — change point detection.
- Learning from data II: Model selection — hypothesis testing.
- Approximate inference: Gibbs sampling and variational Bayes
- Factor models.
- Win 1 million dollars (to share with your supervisor).

Where are the Cancers?

Cancer rate across US counties (kidney cancer rate white males).

The rate for county i is

$$f_i = \frac{n_i}{N_i}$$

Inference

Citation from Dictionary.com

"The process of arriving at some conclusion that, though it is not logically derivable from the assumed premises, possesses some degree of probability relative to the premises."

Setting Up the Model

Write down the probability of the data!

Binomial

$$P(n|\theta, N) = \left(\begin{array}{c} N \\ n \end{array} \right) \theta^n (1-\theta)^{N-n}$$

geometric — probability of sequence $s = (0, 0, 1, 0, 1, 1, 0, \ldots)$

$$P(s|\theta) = \prod_{j=1}^{N} \theta^{s_j} (1-\theta)^{1-s_j} = \theta^n (1-\theta)^{N-n}$$

Learning the Model – Maximum Likelihood

Maximize the probability of the observed data!

$$\widehat{\theta} = \underset{\theta}{\text{argmax}}\, P(s|\theta)$$

$P(s|\theta)$ is also called the likelihood of θ.

Mini student exercise:

1. Find maximum likelihood estimate in geometric distribution.

2. Is the result intuitively reasonable?

3. What about when N is small (for example $N = 1$)?

Learning the Model – Bayes

Enter the prior $P(\theta)$

Bayes theorem

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

Summarizes all we know about θ

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

The Bayesian Cancer Map (handout)

Cancer rate in county i, θ_i is a random variable.

$$\begin{aligned} P(\theta_i) &= \text{Dirichlet}(\theta_i|\alpha,\beta) \\ \text{Dirichlet}(\theta|\alpha,\beta) &= \frac{1}{Z(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ Z(\alpha,\beta) &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \end{aligned}$$

Enter hierarchical modelling:

The cancer rates $\theta_1, \ldots,$ are all drawn from the same distribution.

The parameters are set to match some average properties over the whole US.

Conjugate Distributions

The prior multiplied by the likelihood has the same functional form as the prior

$$\Rightarrow$$

If the prior is easy to handle then the posterior is too.

$$\begin{aligned} P(s|\theta) &= \prod_{j=1}^{N} \theta^{s_j}(1-\theta)^{1-s_j} = \theta^n(1-\theta)^{N-n} \\ P(\theta) &= \frac{1}{Z(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

Mini student exercise

1. Write down the posterior without normalizing it.

2. Sketch prior and posterior for a) $n = 1$, $N = 1$, $\alpha = \beta = 1$ and b) (use computer) $n = 3$, $N = 5$ and $\alpha = \beta = 2$.

3. What is the maximum likelihood estimate in the two cases?

4. What is the maximum value (mode) of the posterior in general?

5. Can we understand the Bayesian cancer map from this?

6. Advanced question: Find the normalization using the normalizer of the prior.

$$P(s|\theta) = \prod_{j=1}^{N} \theta^{s_j}(1-\theta)^{1-s_j} = \theta^n(1-\theta)^{N-n}$$

$$P(\theta) = \frac{1}{Z(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

The Sly Gambler

You are playing a game with a very experienced gambler. In this game it will be advantageous to change the bias of the die used once during the game. You get the feeling that the very experienced gambler has a very suspicious tendency to get the right number of eyes at the right time. You expect that your opponent can change the die at least sometimes without you noticing. You are a scientifically minded person who wants to know whether you are being cheated. What to do?

Exposing the Gambler using Statistical Inference

Let us try to make a series of models for this game:

Special die can only have two outcomes!

Model 1: The gambler only uses one die.

$$P(s|\theta, \mathcal{M}_1) = \theta^n (1-\theta)^{N-n}$$

Model 2: The gambler changes die at throw a:

$$P(s|\theta_1, \theta_2, a, \mathcal{M}_2) = \prod_{j=1}^{a-1} \theta_1^{s_j} (1-\theta_1)^{1-s_j} \prod_{j=a}^{N} \theta_2^{s_j} (1-\theta_2)^{1-s_j}$$

Bayesian inference

Priors:

$$\begin{aligned} P(a|\mathcal{M}_2) &= \frac{1}{N-1} \qquad a \in \{2, 3, \ldots, N\} \\ P(\theta_1|\mathcal{M}_2) &= \text{Dirichlet}(\theta_1|\alpha, \beta) \\ P(\theta_2|\mathcal{M}_2) &= \text{Dirichlet}(\theta_2|\alpha, \beta) \end{aligned}$$

Leave it as an exercise to adjust these priors to include prior information about game (when it is advantageous to change).

Explaining Away — Marginals

Probability of everything

$$P(s, \theta_1, \theta_2, a|\mathcal{M}_2) = P(s|\theta_1, \theta_2, a, \mathcal{M}_2)P(a|\mathcal{M}_2)P(\theta_1|\mathcal{M}_2)P(\theta_2|\mathcal{M}_2)$$

Marginal likelihood = likelihood of model

$$P(s|\mathcal{M}_2) = \int d\theta_1 d\theta_2 \sum_{a=1}^{N} P(s, \theta_1, \theta_2, a|\mathcal{M}_2)$$

Marginal for change point

$$p(a|s, \mathcal{M}_2) = \int d\theta_1 d\theta_2 P(\theta_1, \theta_2, a|s, \mathcal{M}_2)$$

Marginal for bias 1

$$p(\theta_1|s, \mathcal{M}_2) = \int d\theta_2 \sum_{a=1}^{N} P(\theta_1, \theta_2, a|s, \mathcal{M}_2)$$

For details, see Liu & Lawrence, Bioinformatics, 1999.

Model Comparison

Enter prior over models $P(\mathcal{M}_1)$ and $P(\mathcal{M}_2)$

$$\sum_i P(\mathcal{M}_i) = 1$$

Posterior over models

$$P(\mathcal{M}_i|s) = \frac{P(s|\mathcal{M}_i)P(\mathcal{M}_i)}{\sum_{i'} P(s|\mathcal{M}_{i'})P(\mathcal{M}_{i'})}$$

Posterior ratio

$$\frac{P(\mathcal{M}_2|s)}{P(\mathcal{M}_1|s)}$$

If priors are equal: Posterior ratio = Bayes factor

$$\frac{P(\mathcal{M}_2|s)}{P(\mathcal{M}_1|s)} = \frac{P(s|\mathcal{M}_2)P(\mathcal{M}_2)}{P(s)} \frac{P(s)}{P(s|\mathcal{M}_1)P(\mathcal{M}_1)} = \frac{P(s|\mathcal{M}_2)}{P(s|\mathcal{M}_1)} \equiv \mathrm{BF}_{21}$$

Ockham's Razor

Why can we compare models with different number of parameters directly without having to penalize for model complexity (overfitting)?

Because we are averaging and not optimizing.

A flexible model can explain may different data sets $\mathcal{D}$ well

$$\Rightarrow$$

$P(\mathcal{D}|\mathcal{M})$ will be relatively flat.

A simple model can only explain a smaller set of data sets $\mathcal{D}$

$$\Rightarrow$$

$P(\mathcal{D}|\mathcal{M})$ will be more peaked.

Example: regression with different degree polynomials.

Limitations

Cannot detect the absolute quality of a model only the relative within our choice of models.

Extending the example above to model $\mathcal{M}_0$ with no tunable parameters

$$P(s|\theta_0, \mathcal{M}_0) = \theta_0^n (1-\theta_0)^{N-n}$$

and we have no reasons to believe more or less in any of the models

$$P(\mathcal{M}_i) = \frac{1}{3}$$

Posterior changes!

(Human) creativity enters here!

Frequentist Statistics Digression

Is all about the resampling distribution. How will the estimates and likelihoods change if we drew another sample from the sampling distribution $P(s|\theta, \mathcal{M}_i)$

$$\frac{P(s|\widehat{\theta}_1, \widehat{\theta}_2, a, \mathcal{M}_2)}{P(s|\widehat{\theta}, \mathcal{M}_1)}$$

is χ^2 distributed with $3-1=2$ degrees of freedom (nested models).

p-value is the probability of getting an at least as extreme value of this ratio as the observe data set if the null hypothesis $\mathcal{M}_1$ is true.

Not exactly a "man in the street" probability interpretation.

Approximate Inference

We cannot make Bayesian inference exactly for anything but the simplest models

Approximations

- Sampling – MCMC, for example, Gibbs sampling
- Asymptotic result – Bayesian information criterion (BIC)
- Deterministic mean field – Variational Bayes

Markov Chain Monte Carlo

Draw samples from posterior $P(\theta|\mathcal{D})$

$$\theta^{(1)}, \ldots, \theta^{(R)}$$

Approximate average of $f(\theta)$ as

$$\begin{aligned} \langle f(\theta) \rangle &= \int d\theta \, f(\theta) \, P(\theta|\mathcal{D}) \\ &\approx \frac{1}{R} \sum_{r=1}^{R} f(\theta^{(r)}) \end{aligned}$$

Sample $\{\theta^{(r)}\}_{r=1,\ldots,R}$ is called Markov chain because it is generated from $P(\theta^{(r)}|\theta^{(r-1)})$.

Samples correlated $\Rightarrow$ discard samples to get unbiased estimates.

Many other practical issues: convergence of Markov chain (burn-in), step-sizes, etc.

Gibbs Sampling

Just one example of a MCMC method.

Split variables in a number of subsets for example $\theta = \{\theta_1, \theta_2\}$

Many cases impossible to sample from $P(\theta_1, \theta_2|\mathcal{D})$ but easy to sample from conditionals:

$$P(\theta_1|\theta_2, \mathcal{D}) \qquad \text{and} \qquad P(\theta_2|\theta_1, \mathcal{D})$$

Gibbs sampling: Alternate between drawing from each conditional

Equivalent to drawing from posterior

$$P(\theta_1, \theta_2|\mathcal{D}) = P(\theta_1|\theta_2, \mathcal{D})P(\theta_2|\mathcal{D}) = P(\theta_2|\theta_1, \mathcal{D})P(\theta_1|\mathcal{D})$$

Gibbs Sampling – Two Dice

For fixed a, distributions of θ_i, $i = 1, 2$ are simple and vice versa.

- Draw (update) θ_1 for fixed θ_2 and a:

$$P(\theta_1|\theta_2, a, s) = \text{Dirichlet}(\theta_1|\alpha + n_1(a), \beta + a - 1 - n_1(a))$$

  The number of events for die 1: $n_1(a) = \sum_{j=1}^{a-1} s_j$.

- Draw (update) θ_2 for fixed θ_1 and a:

$$P(\theta_2|\theta_1, a, s) = \text{Dirichlet}(\theta_2|\alpha + n_2(a), \beta + N - a - n_2(a))$$

- Draw (update) a for fixed θ_1 and θ_2

$$P(a|\theta_1, \theta_2, s) = \text{Discrete}(a|\theta_1, \theta_2, s)$$

Factor Models

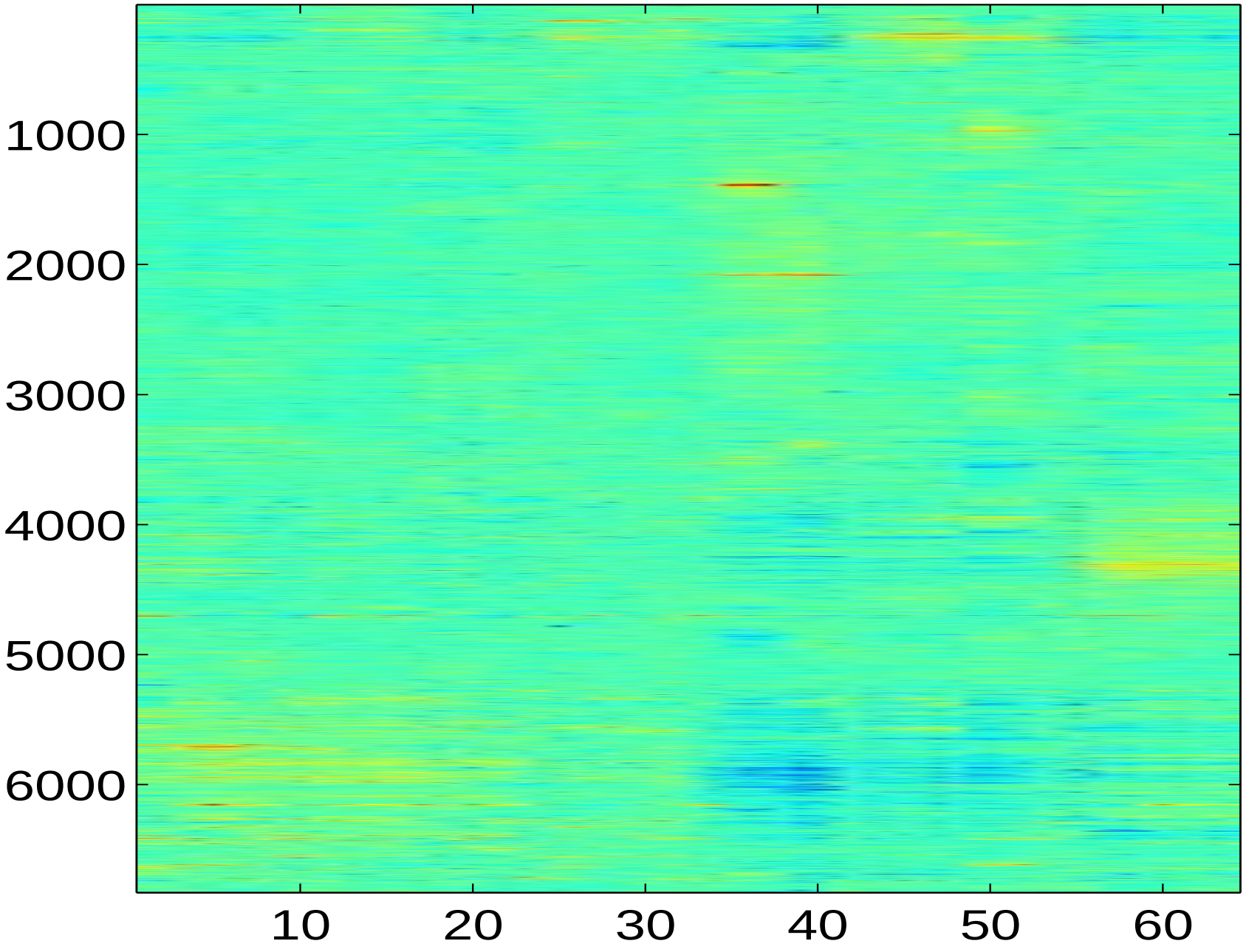
- Source separation (demo!)

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{s}_t + \boldsymbol{\epsilon} \\ \text{recordings}(t) &= \mathbf{A}\ \text{sound sources}(t) + \text{noise} \end{aligned}$$

- Low rank factorization of experimental data (example on next page)

$$\begin{aligned} X_{mn} &= \mathbf{u}_m \cdot \mathbf{v}_n + \epsilon_{mn} \\ \mathbf{X} &= \mathbf{U}\mathbf{V}^T + \mathbf{N} \end{aligned}$$

$\mathbf{U}$ is $M \times K$ and $\mathbf{V}$ is $N \times K$ with $M, N \gg K$.


1000
2000
3000
4000
5000
6000
10
20
30
40
50
60

Mini-exercise - Gibbs Sampler Factor Model

- Joint probability

$$P(\mathbf{X}, \mathbf{U}, \mathbf{V}) = P(\mathbf{X}|\mathbf{U}, \mathbf{V}, \sigma^2)P(\mathbf{U})P(\mathbf{V})$$

- Assume Gaussian prior for $\mathbf{U}$ and $\mathbf{V}$ and iid Gaussian noise

$$P(\mathbf{X}|\mathbf{U}, \mathbf{V}, \sigma^2) = \prod_{m,n} \mathcal{N}(X_{mn}; \mathbf{u}_m \cdot \mathbf{v}_n, \sigma^2)$$

- What type of distribution is

$$P(\mathbf{U}|\mathbf{X}, \mathbf{V}) = \frac{P(\mathbf{X}, \mathbf{U}, \mathbf{V})}{P(\mathbf{X}, \mathbf{V})} \qquad P(\mathbf{X}, \mathbf{V}) = \int P(\mathbf{U}|\mathbf{X}, \mathbf{V}) d\mathbf{U}$$

- We know how to draw samples from a multivariate normal distribution (Bishop 528).

Variational Bayes (VB)

A deterministic method that in many cases has the same complexity as maximum likelihood.

Idea: Introduce a simpler distribution $Q(\theta)$ for which we can carry out all averages.

Example, fully factorized: $Q(\theta) = \prod_i Q_i(\theta_i)$

Minimize difference between $Q(\theta)$ and $P(\theta|\mathcal{D})$.

Kullback-Leibler divergence.

$$KL(Q|P) = \int d\theta\, Q(\theta)\, \log \frac{Q(\theta)}{P(\theta)}$$

General solution for fully factorized

$$Q_i(\theta_i) \propto \exp \langle \log P(\theta, \mathcal{D}) \rangle_{Q \setminus Q_i(\theta_i)} \propto P_i(\theta_i) \exp \langle \log P(\mathcal{D}|\theta) \rangle_{Q \setminus Q_i(\theta_i)}$$

Iterative solution of set of equations for sufficient statistics.

VB – Two Dice

Fully factorized:

$$Q(\theta) = Q(\theta_1)Q(\theta_2)Q(a)$$

Likelihood:

$$P(s|\theta_1, \theta_2, a, \mathcal{M}_2) = \prod_{j=1}^{a-1} \theta_1^{s_j}(1-\theta_1)^{1-s_j} \prod_{j=a}^{N} \theta_2^{s_j}(1-\theta_2)^{1-s_j}$$

VB solution

$$\begin{aligned} Q(\theta_1) &\propto P(\theta_1) \exp\langle \sum_{j=1}^{a-1} \Big[s_j \log\theta_1 + (1-s_j)\log(1-\theta_1) \Big] \rangle_{Q(a)} \\ &= P(\theta_1) \exp \sum_{a=2}^{N} Q(a) \left[n_1(a) \log\theta_1 + (a - n_1(a)) \log(1-\theta_1) \right] \end{aligned}$$

Collecting the factors:

$$Q(\theta_1) = \text{Dirichlet}(\theta_1|\alpha + \langle n_1(a)\rangle, \beta + \langle a\rangle - \langle n_1(a)\rangle)$$

with $\langle n_1(a)\rangle_{Q(a)} = \sum_a n_1(a)\, Q(a)$ and $\langle a\rangle_{Q(a)} = \sum_a a\, Q(a)$.

Mini-exercise: VB Factor Model

- Consider same model as for Gibbs sampler exercise.
- Use factorization

$$Q(\mathbf{U}, \mathbf{V}) = Q(\mathbf{U})Q(\mathbf{V}) \ .$$

- Use general solution

$$Q(\mathbf{U}) = P(\mathbf{U}) \ldots$$

- What type of distribution is $Q(\mathbf{U})$?

Netflix prize

- Netflix - online movie rental (DVDs).
- Collaborative filtering – predict user rating from past behavior of user.
- Improve Netflix own system by 10% to win.
- training.txt – $R = 10^8$ ratings, scale 1 to 5 for $M = 17.770$ movies and $N = 480.189$ users.
- qualifying.txt – 2.817.131 movie-user pairs, (continuous) predictions submitted to Netflix returns a RMSE.
- Rating matrix r_{mn} mostly missing values, 98.5%.

Mini-exercise - Netflix prize

- Discuss simple rules to make predictions.
- Sketch machine learning approaches.
- What prior knowledge can we use?
- How to handle discrete categories?
- Is the size of the data set an issue?
- Sign up for Netflix 02459 project to try it out.

Summary

- Bayes - use a priori information.
- Approximate inference - Gibbs sampling and variational Bayes.
- Factor models.
- Netflix prize.