C19 : Lecture 2 : Inference, Learning, Monte Carlo Integration, Basic Sampling

Frank Wood

University of Oxford

January, 2017

Many figures from PRML [Bishop, 2006]

Inference

In probabilistic modelling a model corresponds to some specification of the distribution of either the data given parameters (supervised) $p(\mathcal{D}|\mathcal{M})$ or the joint distribution of the data and parameters $p(\mathcal{D}, \mathcal{M}) = p(\mathcal{D}|\mathcal{M})p(\mathcal{M})$ (unsupervised).

In the the latter, learning, inference, and prediction can be uniformly cast as "simple" probability calculations.

> *Bayes' rule tells us that the posterior probability of a model $\mathcal{M}$ given a set of observations $\mathcal{D}$ is proportional to the likelihood of the obervations under the model $p(\mathcal{D}|\mathcal{M})$ times the prior probability of the model $\mathcal{M}$, $p(\mathcal{M})$*

Useful Calculations Using The Posterior / Joint

Learning

Maximum A Posteriori Model Selection

$$\mathcal{M}^* = \underset{\mathcal{M}}{\text{argmax}}\, p(\mathcal{M}|\mathcal{D}) = \underset{\mathcal{M}}{\text{argmax}}\, p(\mathcal{D}|\mathcal{M})p(\mathcal{M})$$

Model Selection Via Bayes Factors

...

Prediction

Posterior Predictive

$$p(d'|\mathcal{D}) = \int p(d'|\mathcal{M})p(\mathcal{M}|\mathcal{D})d\mathcal{M}$$

Inspection

If model family is parameterised by a single parameter $\theta \in \mathbb{R}$ then

$$p(a \leq \theta \leq b|\mathcal{D}) \equiv \int p(\theta|D)\mathbb{I}(a \leq \theta \leq b)d\theta$$

Integration

Integration/marginalisation is in general hard

- Simple analytic functions
  - High school maths
- Low-dimensional, well-behaved functions
  - Quadrature
- Nice functions
  - Dynamic programming, sum-product
- Everything else
  - Analytic Integration
    - *Important :* Bayesian statistics, particularly in the exponential family exploiting conjugacy
  - Approximation
    - *Important :* Sampling (MCMC, SMC)
  - Learning and integrating against simple, surrogate approximating functions
    - Variational inference
    - Bayesian quadrature

Analytic Integration Integration Example

With

$$\begin{aligned} \theta &\sim \text{Dir}_K(\alpha\vec{1}) \\ z_i &\sim \text{Discrete}(\theta) \end{aligned}$$

what is $p(z_{i+1}|\{z_i\}_{i=1}^N)$?

Recall that if $\theta \sim \text{Dir}_K(\alpha)$ then

$$p(\theta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_K^{\alpha_K - 1}.$$

Analytic Integration Integration Example

Note

$$p(z_{i+1}|\{z_i\}_{i=1}^N) = \int p(z_{i+1}|\theta)p(\theta|\{z_i\}_{i=1}^N)d\theta$$

First let $N_k = \sum_{i=1}^N \mathbb{I}(z_i = k)$ and note that

$$\begin{aligned} p(\theta|\{z_i\}_{i=1}^N) &\propto p(\{z_i\}_{i=1}^N|\theta)p(\theta) \\ &\propto \left(\prod_k \theta_k^{N_k}\right) \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\ldots\theta_K^{\alpha_K-1} \end{aligned}$$

$$\implies \theta|\{z_i\}_{i=1}^N \sim \mathrm{Dir}_K(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

Analytic Integration Integration Example

So analytic integration with Dirichlet normalisation solves the integral

$$
\begin{aligned}
& p(z_{i+1} = j | \{z_i\}_{i=1}^N) \\
&= \int p(z_{i+1}|\theta) p(\theta|\{z_i\}_{i=1}^N) d\theta \\
&= \int \theta_{z_{i+1}} \frac{\Gamma(\sum_i (N_i + \alpha_i))}{\prod_i \Gamma(N_i + \alpha_i)} \theta_1^{\alpha_1 + N_1 - 1} \dots \theta_K^{\alpha_K + N_K - 1} d\theta \\
&= \frac{\Gamma(\sum_i (N_i + \alpha_i))}{\prod_i \Gamma(N_i + \alpha_i)} \int \theta_1^{\alpha_1 + N_1 - 1} \dots \theta_j^{\alpha_j + N_j + 1 - 1} \dots \theta_K^{\alpha_K + N_K - 1} d\theta
\end{aligned}
$$

$$
\implies p(z_{i+1} = j | \{z_i\}_{i=1}^N) = \frac{\Gamma(\sum_i (N_i + \alpha_i))}{\prod_i \Gamma(N_i + \alpha_i)} \frac{\prod_i \Gamma(N_i + \mathbb{I}(i = j) + \alpha_i)}{\Gamma(\sum_i (N_i + \mathbb{I}(i = j) + \alpha_i))}
$$

Homework : Using facts about Γ function, this simplifies dramatically.

Interpretation of Probability Computations

Example : Let θ_k represent the proportion of customers of age between $10k$ and $10(k+1)$ with $K = 15$ that are served by an organisation. Then

- α is the "prior" belief about what the proportions should be
- z_i is the age bracket of the ith observed customer
- $\theta|\{z_i\}_{i=1}^N$ is the "learned" customer age distribution distribution, represented with uncertainty (variance) that shrinks as $N \to \infty$
- $z_{i+1}|\{z_i\}_{i=1}^N$ allows you to predict the next customer's age whilst taking into account uncertainty about the value of θ

Important

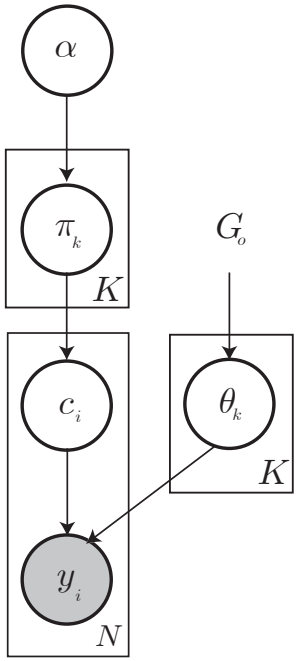
Learning, prediction, and inspection can all be expressed in terms of analytic and computational manipulation of probability (marginalisation and conditioning) within a model specified as a joint distribution of data and parameters.

Harder Problem

Assume that we would like to know whether or not two "instances" are of the same class, or, equivalently, are "similar"

- Unsupervised problem – no class+feature training data, only features
- Quantity of interest not directly observable (called latent or hidden)
- Requires explicit model

Introduction : Gaussian Mixture Model



$$
\begin{aligned}
c_i|\vec{\pi} &\sim \text{Discrete}(\vec{\pi}) \\
\vec{y}_i|c_i = k; \Theta &\sim \text{Gaussian}(\cdot|\theta_k). \\
\vec{\pi}|\alpha &\sim \text{Dirichlet}(\cdot|\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}) \\
\Theta &\sim \mathcal{G}_0
\end{aligned}
$$

Joint

$$
\begin{aligned}
& p(\mathcal{Y}, \Theta, \mathcal{C}, \vec{\pi}, \alpha; \mathcal{G}_0) \\
&= \left(\prod_{k=1}^{K} p(\theta_k; \mathcal{G}_0) \right) \left(\prod_{i=1}^{N} p(\vec{y}_i|c_i, \theta_{c_i}) p(c_i|\vec{\pi}) \right) P(\vec{\pi}|\alpha) p(\alpha)
\end{aligned}
$$

Mapping from Problem to Model

Assume that we would like to know whether or not two "instances" are of the same class, or, equivalently, are "similar"

- Features of instances $\vec{y}_i$
- Latent classes c_i
- Characterised by per-class generative models parameterised by θ_k

Solution as integration

$$p(c_i = c_j|\mathcal{Y}) \propto \int\int\int\int\int p(\mathcal{Y}, \Theta, \mathcal{C}, \vec{\pi}, \alpha; \mathcal{G}_0)\mathbb{I}(c_i = c_j)d\Theta d\mathcal{C} d\vec{\pi} d\alpha$$

What Looks Horrible Often Is Not

While

$$p(c_i = c_j|\mathcal{Y}) \propto \int\int\int\int\int p(\mathcal{Y}, \Theta, \mathcal{C}, \vec{\pi}, \alpha; \mathcal{G}_0)\mathbb{I}(c_i = c_j)d\Theta d\mathcal{C} d\vec{\pi} d\alpha$$

looks horrible, it is actually rather easy.

Homework:

- Show that you can analytically integrate out everything except the c's
- What choices for $\mathcal{G}_0$ make this possible?
- Why can't the sum over the c's be performed?

Aside : How to Build a Model for Your Own Problem

From where do models like this come?

Statistics, machine learning, and other fields consist of the exploration and characterisation the space of models. Most models you see are "convenient" in that much integration is analytically possible and that inference, inspection, and prediction are all known to work well and reliably for a wide variety of problems. The design of new models involves equal parts knowledge of a problem domain, familiarity with statistical model building blocks, and mathematical/computational aesthetic. Often it is possible to creatively map your specific inference problem onto an existing model saving time and effort.

Compressed Notation for Sampling

Generalizing

$$p(c_i = c_j|\mathcal{Y}) \propto \int\int\int\int\int p(\mathcal{Y}, \Theta, \mathcal{C}, \vec{\pi}, \alpha; \mathcal{G}_0)\mathbb{I}(c_i = c_j)d\Theta d\mathcal{C} d\vec{\pi} d\alpha$$

Inference, prediction, and inspection can all be expressed as expectations

$$\mathbb{E}[f] \equiv \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

Where $\mathbf{x}$ is *all* latent variables, f is a test function, and p is the distribution against which we're integrating.

A Don't Fix it If it Ain't Broke Tool : Monte Carlo Integration

Recipe

1. Sample $\mathbf{x}^{(\ell)} \sim p(\mathbf{x})$ for $\ell = 1 \dots L$
2. Estimate $\mathbb{E}[f] \approx \hat{f} = \frac{1}{L} \sum_{\ell=1}^{L} f(\mathbf{x}^{(\ell)})$

Claim (1) : $\hat{f}$ is unbiased, i.e. $\mathbb{E}[\hat{f}] = \mathbb{E}[f]$

$$
\begin{aligned}
\mathbb{E}[\hat{f}] &= \mathbb{E}\left[\frac{1}{L}\sum_{\ell=1}^{L} f(\mathbf{x}^{(\ell)})\right] = \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}[f(\mathbf{x}^{(\ell)})] \\
&= \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{E}[f(\mathbf{x})] \qquad \mathbb{E}[f(\mathbf{x}^{(j)})] = \mathbb{E}[f(\mathbf{x}^{(k)})] \text{ since } \mathbf{x}^{(\ell)} \text{ iid} \\
&= \mathbb{E}[f] \quad \square
\end{aligned}
$$

Claim (2) : The variance of $\hat{f}$ is independent of the dimensionality of $\mathbf{x}$ and decreases at a rate of $1/L$

$$\begin{aligned}
\mathrm{Var}[\hat{f}] &= \mathrm{Var}\left[\frac{1}{L}\sum_{\ell=1}^{L} f(\mathbf{x}^{(\ell)})\right] = \frac{1}{L^2}\sum_{\ell=1}^{L}\mathrm{Var}[f(\mathbf{x}^{(\ell)})] \qquad \text{since } \mathbf{x}^{(\ell)} \text{ iid} \\
&= \frac{1}{L}\sum_{\ell=1}^{L}\mathrm{Var}[f(\mathbf{x})] \qquad \text{since } \mathbf{x}^{(\ell)} \text{ iid} \\
&= \frac{1}{L}\mathbb{E}[(f-\mathbb{E}[f])^2] \quad \square
\end{aligned}$$

If f is poorly behaved, L is small, or the $\mathbf{x}^{(\ell)}$s are correlated, the variance of the estimator could be quite large. In sampling-based inference there is usually a difficult practical trade-off between the latter two.

It All Boils Down To Sampling

Monte Carlo integral approximation boils down to designing, characterising, and running samplers.

Types of sampling we will cover include

- Rejection Sampling
- Conditioning via Rejection and Ancestral Sampling
- Metropolis Hastings
- Gibbs Sampling
- Importance Sampling
- Sequential Monte Carlo

Types of sampling we will *not* cover include

- Random number generation, variable transformations, sampling from standard distributions

Rejection Sampling I

Assume

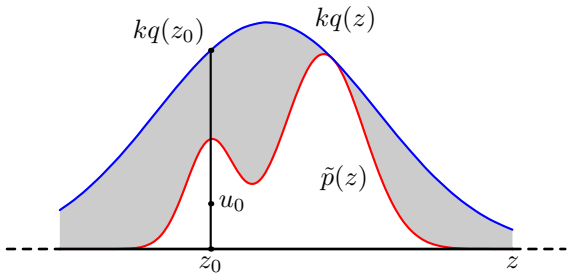
- Want sample from $p(\mathbf{x})$
- $p(\mathbf{x})$ is easy to evaluate, but only up to an unknown normalising constant, i.e.

$$p(\mathbf{x}) = \frac{1}{Z_p}\tilde{p}(\mathbf{x})$$

- A *proposal distribution* $q(\mathbf{x})$ s.t. $kq(\mathbf{x}) \geq \tilde{p}(\mathbf{x})$ for all $\mathbf{x}$ can be designed

Note $\mathbf{x}$ is, in general, a vector of random variables.

Rejection Sampling II



Sampling $\mathbf{x}^{(\tau)} \sim q$ and $u^{(\tau)} \sim \text{Uniform}(0, kq(\mathbf{x}^{(\tau)}))$ yields a pair of values uniformly distributed in the gray region.

If $u_0 \leq \tilde{p}(\mathbf{x})$ then $\mathbf{x}^{(\tau)}$ is accepted, otherwise it is rejected and the process repeats until a sample is accepted.

Accepted pairs are uniformly distributed in the white area; dropping $u^{(\tau)}$ yields a sample distributed according to $\tilde{p}(\mathbf{x})$, and equivalently, $p(\mathbf{x})$.

The efficiency of rejection sampling depends critically on the match between the proposal distribution and the distribution of interest.

Conditioning via Rejection and Ancestral Sampling I

Assume we have a model $p(\mathbf{x})$, some variables of which are known, some of which are not. Also let $\mathbf{x}_{\text{obs}}$ be the "observed" variables and $\mathbf{x}_{\text{lat}}$ be latent variables such that $\mathbf{x}_{\text{obs}} \cup \mathbf{x}_{\text{lat}} = \mathbf{x}$.

We would like samples from $p(\mathbf{x}_{\text{lat}}|\mathbf{x}_{\text{obs}}) = \frac{p(\mathbf{x})}{p(\mathbf{x}_{\text{obs}})}$

Equivalently we can write the conditional distribution of interest as an unnormalised distribution $\tilde{p}(\mathbf{x}_{\text{lat}}|\mathbf{x}_{\text{obs}}) = p(\mathbf{x})\mathbb{I}[\mathbf{x}_{\text{obs}} = \mathbf{v}]$ using an indicator function that imposes the constraint that the observed variables are constrained to take values $\mathbf{v}$.

Rejection sampling with $q(\mathbf{x}) = p(\mathbf{x})$ (i.e. proposing via ancestral sampling of the joint) can be used to generate samples distributed according to $\tilde{p}(\mathbf{x}_{\text{lat}}|\mathbf{x}_{\text{obs}})$. Note that $q(\mathbf{x}) \geq \tilde{p}(\mathbf{x}_{\text{lat}}|\mathbf{x}_{\text{obs}})\ \forall \mathbf{x}$ by construction.

Conditioning via Rejection and Ancestral Sampling II

Following the rejection sampling recipe yields a posterior conditional sampler via ancestral sampling and rejection

Conditioning via Rejection and Ancestral Sampling

1. Sample $\mathbf{x}^{(\tau)} \sim q(\mathbf{x})$ (i.e. generate via ancestral sampling)
2. Sample $u^{(\tau)} \sim \mathrm{U}(0, q(\mathbf{x}))$
3. Accept $\mathbf{x}^{(\tau)}$ only if $u^{(\tau)} \leq p(\mathbf{x})\mathbb{I}[\mathbf{x}_{\mathrm{obs}} = \mathbf{v}]$
4. Repeat

A sample will only ever be accepted when $\mathbf{x}_{\mathrm{obs}} = \mathbf{v}$ and then it will always be because $q(\mathbf{x}) = p(\mathbf{x})$

Unless the prior and posterior are extremely well matched this will be an extremely inefficient sampler.

Study Suggestions

- Conjugacy
- Exponential family distributions
- Bayes rule
- Theory of statistical estimators
- Sampling from common distributions
- Practice phrasing an analysis in terms of an expectation

Bibliography I

C M Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.