Bayesian Learning

Lecture 11 - Course summary

Mattias Villani

Department of Statistics
Stockholm University

mattiasvillani.com @matvil @matvil mattiasvillani

Overview

- **Wait, what did we actually do?**

Bayesics

- **Subjective probability**
- **Bayesian Learning**: Bayes theorem

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|\theta)p(\theta)$$

- The proportional constant is actually the **marginal likelihood**

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)p(\theta)d\theta$$

used for model comparison.

Simple models - conjugate priors

- iid **Bernoulli model** - Beta prior
- iid **Poisson model** - Gamma prior
- iid **Exponential model** - Gamma prior
- iid **Normal known variance** - Normal prior

Multi-parameter models

- **Marginalization**

$$p(\theta_1|\boldsymbol{x}) = \int p(\theta_1, \theta_2|\boldsymbol{x}) d\theta_2$$

- iid **normal with unknown variance** $N(\theta, \sigma^2)$. Marginal posterior for θ is student-t

$$\theta|\boldsymbol{x} \sim t\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}, \nu_n\right)$$

- **Multinomial model** - Dirichlet prior
- **Dirichlet distribution** on unit simplex $\sum_{k=1}^{K} \theta_k = 1$.

Linear Gaussian regression

- **Linear regression**

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- **Conjugate prior**

$$\boldsymbol{\beta}|\sigma^2 \sim N\left(\mu_0, \sigma^2 \Omega_0^{-1}\right)$$
$$\sigma^2 \sim \text{Inv} - \chi^2\left(\nu_0, \sigma_0^2\right)$$

- **Posterior**

$$\boldsymbol{\beta}|\sigma^2, \mathbf{y} \sim N\left(\mu_n, \sigma^2 \Omega_n^{-1}\right)$$
$$\sigma^2|\mathbf{y} \sim \text{Inv}{-}\chi^2\left(\nu_n, \sigma_n^2\right)$$

- **L2-Regularization** prior (Ridge)

$$\boldsymbol{\beta}|\sigma^2 \sim N\left(\mathbf{0}, \frac{\sigma^2}{\lambda} \boldsymbol{I}_p\right)$$

- **L1-Regularization** prior (Lasso)

$$\boldsymbol{\beta}|\sigma^2 \sim \text{Laplace}\left(\mathbf{0}, \frac{\sigma^2}{\lambda} \boldsymbol{I}_p\right)$$

Logistic regression and Beyond

- **Logistic regression**

$$\Pr(y_i = 1|\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

- Posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X})$ is intractable. Solutions:
  - Normal approximation
  - Posterior sampling using MH/HMC
- Linear Gaussian regression is modeling a conditional density

$$y_i|\boldsymbol{x}_i \overset{\text{ind}}{\sim} N(\mu_i, \sigma^2) \text{ where } \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

- Paves the way for **Poisson regression**:

$$y_i|\boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Pois}(\mu_i) \text{ where } \mu_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$$

- **Exponential regression**

$$y_i|\boldsymbol{x}_i \overset{\text{ind}}{\sim} \text{Expon}(\theta_i) \text{ where } \theta_i = \frac{1}{e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}$$

so that $\mathbb{E}(y_i|\boldsymbol{x}_i) = \mu_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$.

Priors

- **Expert** elicitation
- **Other data** sources
- **Non-informative priors** (prior sample size). Bernoulli model:

$$\theta|\boldsymbol{x} \sim \text{Beta}(\alpha + s, \beta + f)$$

- **Jeffreys' prior** $p(\theta) = |I(\theta)|^{1/2}$
- **Hierarchical priors**. L2-regularization

$$p(\boldsymbol{\beta}, \sigma^2, \lambda) = p(\boldsymbol{\beta}|\sigma^2\lambda)p(\sigma^2)p(\lambda)$$

- **Regularization** and **smoothness priors**

Prediction

- **(Posterior) predictive distribution** for iid data:

$$p(\tilde{y}|\boldsymbol{y}) = \int_{\theta} p(\tilde{y}|\theta)p(\theta|\boldsymbol{y})d\theta$$

  - $p(\tilde{y}|\theta)$ is the model density for a new observation $\tilde{y}$ given θ
  - $p(\theta|\boldsymbol{y})$ is the posterior density for θ given the training data $\boldsymbol{y}$

- With no training data: the **prior predictive distribution**

$$p(\tilde{y}) = \int_{\theta} p(\tilde{y}|\theta)p(\theta)d\theta$$

  where we integrate of the parameters using the **prior** $p(\theta)$.

- Use prior predictive density to set the prior hyperparameters when the expert expresses prior beliefs about observable data:
  - $\mathbb{E}(\tilde{y})$
  - $\Pr(\tilde{y} > c)$

- Interactive example with Gamma prior for Poisson model.

Decisions

- Need to make a **decision/action** $a \in \mathcal{A}$ when **state of nature** $\boldsymbol{\theta}$ is partially unknown.
- The **utility function** describes the consequence of taking action a when state of nature is $\boldsymbol{\theta}$

$$U(a, \boldsymbol{\theta})$$

- The utility function is subjective (also for non-Bayesians) and is determined by the decision maker.
- **Bayesian theory** gives a single universal decision rule: choose the action that **maximizes (posterior) expected utility**:

$$\mathrm{EU}(a) = \int U(a, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}$$

where $\boldsymbol{y}$ is some training data.

Posterior approximation

- Three ways to approximate the posterior:
  - **Normal approximation**
  - **Posterior sampling** (Gibbs, MH, HMC)
  - **Numerical integration**
- Normal approximation is fast, but is approximate unless we have a lot of data (infinite in theory).
- Posterior sampling will converge to true posterior if simulated long enough, but is approximate in finite time.
- Numerical integration is only a good option when the number of parameters is small.

Model evaluation and comparison

- Three main ways to compare models
  - **Posterior model probabilities** (**marginal likelihood**)
  - **Log Predictive Density Score** (LPDS/elpd)
  - **Leave-one-out (LOO)** LPDS/elpd
- Evaluate the **predictive density performance** on test data

$$p(\boldsymbol{y}_{\text{test}}|\boldsymbol{y}_{\text{train}}) = \int p(\boldsymbol{y}_{\text{test}}|\boldsymbol{\theta}, \boldsymbol{y}_{\text{train}}) p(\boldsymbol{\theta}|\boldsymbol{y}_{\text{train}}) d\boldsymbol{\theta}$$

  - test data $\boldsymbol{y}_{\text{test}}$
  - posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_{\text{train}})$ based on some training data.
- Often evaluated on the log scale: $\log p(\boldsymbol{y}_{\text{test}}|\boldsymbol{y}_{\text{train}})$.
- **Parameter uncertainty is included** in the model comparison by integrating with respect to the posterior $p(\boldsymbol{\theta}|\boldsymbol{y}_{\text{train}})$.

Model evaluation and comparison

- **Marginal likelihood**:
  - no $\boldsymbol{y}_{\text{train}}$ and $\boldsymbol{y}_{\text{test}}$ is all the data.
  - Sensitive to the prior on the parameters $p(\boldsymbol{\theta})$.
- **Log Predictive Density Score**:
  - $\boldsymbol{y}_{\text{train}}$ and $\boldsymbol{y}_{\text{test}}$ is some partition of the data.
  - K-fold cross-validation.
- **Leave-one-out (LOO)**
  - $\boldsymbol{y}_{\text{test}} = y_i$ and $\boldsymbol{y}_{\text{train}} = \boldsymbol{y}_{-i}$ (all data except obs i)
  - n-fold cross-validation.
  - Time-consuming: posterior sampling from n posteriors

$$p(\boldsymbol{\theta}|\boldsymbol{y}_{-i}) \text{ for } i = 1, \ldots, n$$

  - The **loo** package uses a importance sampling trick for speed.