Bayesian prediction and model selection

Bayesian Computation 2 / 14

Guillaume Dehaene

Objectives

In this first course, we will continue to explain how to solve the three basic questions of statistics in the Bayesian paradigm

1. Prediction
2. Model selection

We will also start to see that computing the posterior is tricky.

Bayesian prediction

The importance of prediction

Two very important role for statistics:

- Predict the future.
- Reveal the unseen.

How? Learn the correlation between:

- Easy to access predictor variables $X_1 \dots X_d$.
- Hard to access variable of interest Y.

Predicting future sales

Alice wants to sell ice-cream at Ouchy during the summer.
She needs to accurately forecast how much ice-cream she will sell during the day.

Assume that over the last week, she has sold the following quantities (in L):

```
Y = np.array( [40, 60, 55, 75, 80, 75, 85], dtype = np.int
```

She was never sold out (yet).

How can we try to predict the quantity of ice cream she might sell tomorrow Y_{n+1}?

Two key ideas:

- The answer needs to be encoded into a probability distribution:

$$Y \sim F_Y = ??$$

- We need to derive this probability distribution through the application of Bayes' rule.

Ideas?

Predicting future sales

Two key ideas:

- The answer needs to be encoded into a probability distribution:

$$Y \sim F_Y = ??$$

- We need to derive this probability distribution through the application of Bayes' rule.

Ideas?

Solution: **augment** the model with **another variable** !

Predicting future sales

Normal model:

- Conditional model:
$$f\left(Y_1 \ldots Y_7 | \boldsymbol{\theta}\right)$$
- Prior model $f\left(\boldsymbol{\theta}\right)$

Augmented model:

- Add a conditional model for Y_{n+1}.

Predicting future sales

For example:

- Assume that all quantities are IID Gaussian:

$$Y_i \overset{IId}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$$

  includes the observed $Y_1 \dots Y_7$ and the unobserved Y_{n+1}.
- Two unknowns: μ, σ^2.
  - For technical reasons, we will work instead with the precision $\beta = \sigma^{-2}$.
  - For technical reasons, we consider the following prior:

$$\beta \sim \Gamma\left(a = 1, b = 100\right)$$
$$\mu|\beta \sim \mathcal{N}\left(50, \beta^{-1} * 1\right)$$

  (A Gamma-normal hierarchical distribution)

Predicting future sales

Prior density:

$$
\begin{aligned}
f(\beta,\mu) &\propto \beta^{1/2} \exp\left(-100\ \beta - \frac{\beta}{2}(\mu-50)^2\right) \\
&\propto \beta^{a-1/2} \exp\left(-b\ \beta - \frac{\beta}{2}(\mu-50)^2\right)
\end{aligned}
$$

Likelihood:

$$
\begin{aligned}
f(y_1 \dots y_7|\mu,\beta) &\propto \prod_{i=1}^{7} \beta^{1/2} \exp\left(-\frac{\beta}{2}(y_i-\mu)^2\right) \\
&\propto \beta^{7/2} \exp\left(-\frac{\beta}{2}\sum_{i=1}^{7}(y_i-\mu)^2\right) \\
&\propto \beta^{7/2} \exp\left(-\frac{\beta}{2}\left\{7(\bar{y}-\mu)^2 + \sum_{i=1}^{7}(y_i-\bar{y})^2\right\}\right) \\
&\propto \beta^{7/2} \exp\left(-\frac{7(\bar{y}-\mu)^2}{2}\ \beta - \frac{\beta}{2}\sum_{i=1}^{7}(y_i-\bar{y})^2\right)
\end{aligned}
$$

Posterior:

$$f(\beta,\mu|d) \propto \beta^{8/2} \exp\left(-\left\{100 + \frac{\sum_{i=1}^{7}(y_i-\bar{y})^2}{2}\right\}\beta - \frac{\beta}{2}\left\{(\mu-50)^2 + 7(\bar{y}\right.\right.$$

We recognize another Gamma-Normal distribution (!!?!). Defining:

$$\hat{\mu} = \frac{50+7\bar{y}}{1+7}$$

- $\beta|d$ is marginally Gamma:

$$\beta|d \sim \Gamma\left(a = 1 + \frac{n}{2}, b = 100 + \frac{\sum_{i=1}^{7}(y_i-\hat{\mu})^2}{2} + \frac{1}{2}(\hat{\mu}-50)^2\right)$$

- While $\mu|\beta,d$ is Gaussian:

$$\mu|\beta,d \sim \mathcal{N}\left(\hat{\mu},(8\beta)^{-1}\right)$$

Now, we can compute the posterior of the new observation Y_{n+1}. Conditional on μ, β, d, it is Gaussian:

$$Y_{n+1}|\mu, \beta, d \sim \mathcal{N}\left(\mu, \beta^{-1}\right)$$

Marginalizing out μ:

$$\begin{aligned} Y_{n+1}|\beta, d &\sim \mathcal{N}\left(\hat{\mu}, \beta^{-1} + (8\beta)^{-1}\right) \\ &\sim \mathcal{N}\left(\hat{\mu}, \frac{9}{8}\beta^{-1}\right) \end{aligned}$$

Marginalizing out β is harder: it gives a student distribution:

$$Y_{n+1}|d \sim T\left(\text{ddof} = n+2, \mathbb{E} = \hat{\mu}, \sigma^2 = \frac{100 + \frac{\sum_{i=1}^{7}(y_i - \hat{\mu})^2}{2} + \frac{1}{2}\left(\hat{\mu} - 50\right)^2}{8/9\,(n+2)}\right)$$

(Don't sue me if I got it wrong)

Take home messages

Keys:

- Bayesian inference can involve quite a bit of work.
- Importantly, here, the normalization constant did not matter
- Prediction involves the **addition of more variables in the model**.
- We obtain a posterior over the variable to be predicted:

$$f\left(Y_{n+1}|d\right)$$

  We can then construct normal Bayesian point estimates:
  - mean, median, MAP

  or interval estimates:
  - Credible intervals

Take home messages

- Magical coincidence: we recovered a posterior inside the same family as the prior.
  - This is called a conjugate family associated to a conditional model.
  - **This is extremely rare**.
    I chose this feature on purpose to make my life simple.
  - We'll discuss this more next week.
- Prior is partially interpretable since it plays a role comparable to the data ("Pseudo-data" interpretation):
  - For example, the prior mean ($\mathbb{E}(\mu) = 50$) and the empirical mean $\bar{y}$ play the same role in the final formula.
- This is again a property of conjugate families that we will talk about next week.

Model selection

Choosing the right model

In many situations, a number of qualitatively different models could explain the data. The job of the statistician then consists in determining which one is the best.

- Dependance or independance of two measured variables
- Is situation A different or identical to situation B.
- Which predictors are useful for anticipating the value of Y.

This problem of model selection is probably **the hardest problem of statistics.**

Choosing the right model

Classical approaches:

- Neyman-Pearson:
  - Heavily biased towards scientific inference.
  - Two alternatives: H_0 and H_1.
  - Asymetric: reject or conserve H_0.
- Best validation Performance:
  - For each model, find the best fit on a training data set.
  - Compute the "performance" of the best fit on a new data set.
  - Optional: correct for the number of parameters (AIC, BIC, etc)
  - Choose the model with the best validation performance.

Optimizing ice-cream sales

Let's return to Alice and here ice-creams.

Assume she wants to know whether doing something different (e.g. changing the price of the ice-cream, or the recipe) modifies the amount of money she makes in a day.

First, we need to collect data. Let's assume that:

- Each day, **she chooses randomly** whether she will in condition 1 or 2.
- She has collected n observations from each case: X_i and Y_i.
- Does the intervention matter?

We want to know whether intervention matters or not:

- Once again, the answer needs to be a probability distribution.
- That is derived through applying Bayes' rule.

Ideas?

We want to know whether intervention matters or not:

- Once again, the answer needs to be a probability distribution.
- That is derived through applying Bayes' rule.

Ideas?

Once again, the solution consists in augmenting the model with more variables.

A key idea of Bayesian model choice: sampling from the prior should generate realistic datasets (generative approach to priors).

- NB: sampling from the prior means:
    - Choosing a random $\boldsymbol{\theta}$ from the prior.
    - Choosing a random dataset from the conditional distribution $f\left(\mathcal{D}=d|\boldsymbol{\theta}\right)$.

Here, sampling from the prior should generate:

- Some datasets for which the intervention does nothing.
- Some datasets for which the intervention does something.

Optimizing ice-cream sales

An elegant solution for this: the addition of a “Flag” variable: a discrete variable $F \in \{0, 1\}$.

- $F = 1$ corresponds to the active model: intervention does something.
- $F = 0$ corresponds to the inactive model: intervention does nothing.

Optimizing ice-cream sales

Prior distribution:

- Sample $F \sim B(p)$.
- Conditional on $F = 0$

$$\mu_X = \mu_Y \sim \mathcal{N}\left(300, (50)^2\right)$$

.

- Conditional on $F = 1$

$$\mu_X, \mu_Y \overset{IID}{\sim} \mathcal{N}\left(300, (50)^2\right)$$

- Conditional on μ_X, μ_Y:

$$X_i \overset{IID}{\sim} \mathcal{N}(\mu_X, 1)$$
$$Y_i \overset{IID}{\sim} \mathcal{N}(\mu_Y, 1)$$

Optimizing ice-cream sales

We have defined the model. Now comes the painful part where we apply Bayes' rule.

We know how to perform the inference conditional on the value of F: it's just simple inference for a Gaussian model. We thus have:

$$f\left(\mu_X, \mu_Y | F = 0, d\right)$$
$$f\left(\mu_X, \mu_Y | F = 1, d\right)$$

In order to finish, we only need to characterize the marginal distribution: $f\left(F|d\right)$.
Applying Bayes' rule to the pair F, d yields:

$$f\left(F|d\right) \propto f\left(F\right) f\left(d|F\right)$$

where $f\left(d|F\right)$ is the distribution of d when I marginalize out μ_X, μ_Y.

$$f\left(d|F\right) = \int f(d \;\&\; \mu_X, \mu_Y | F) d\mu_X d\mu_Y$$

Optimizing ice-cream sales

This term is precisely the normalizing constant that we obtain when we apply Bayes' rule conditional on the value of F to compute the posterior of μ_X, μ_Y:

$$f(\mu_X, \mu_Y | F, d) = \frac{f(\mu_X, \mu_Y | F) f(d | \mu_X, \mu_Y, F)}{f(d|F)}$$

This is why and where the normalizing constant matters in Bayesian inference: in order to perform model selection !!!

Thus, the overall logic of Bayesian model selection is the following:

- First, perform inference in each model, i.e. conditional on $F = 0$ or $F = 1$.

$$f(\mu_X, \mu_Y | F, d) = \dots$$

  Critically, we need to evalute the normalization constant $f(d|F)$!!

- Then, perform inference for the "Flag" variable:

$$f(F|d) \propto f(F) f(d|F)$$

  Here, the normalization constant does not matter.

Optimizing ice-cream sales

Thankfully, I've chosen a simple model for which the calculation of $f(d|F)$ is simpler.

For $F = 0$:

$$\mu = \mu_X = \mu_Y \sim \mathcal{N}\left(300, (50)^2\right)$$
$$X, Y \overset{IID}{\sim} \mathcal{N}(\mu, 1)$$

Thus, the marginal distribution of X, Y is a Gaussian with parameters:

$$\mathbb{E}(X_i) = \mathbb{E}(Y_i) = 300$$
$$\mathrm{Var}(X_i) = \mathrm{Var}(Y_i) = 1 + 2500$$
$$\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_i, Y_j) = 2500$$

Thus:

$$f(d|F=0) = \frac{(2\pi)^{2n/2}}{|\mathrm{Cov}|^{1/2}} \exp\left(-\frac{1}{2}([X,Y]-300)(\mathrm{Cov})^{-1}([X,Y]-300)\right)$$

The same logic applies for $F = 1$ except the covariance matrix is slightly different:

$$\mathrm{Cov}\left(X_i, Y_j\right) = 0$$

Once again:

$$f\left(d|F=0\right) = \frac{\left(2\pi\right)^{2n/2}}{\left|\mathrm{Cov}\right|^{1/2}} \exp\left(-\frac{1}{2}\left(\left[X,Y\right]-300\right)\left(\mathrm{Cov}\right)^{-1}\left(\left[X,Y\right]-300\right)\right)$$

Take home messages

Keys:

- Once again, notice how much work we had to do on the posterior.
- Here, the normalization constant **of the intermediate variables** μ_X, μ_Y played a key role.
- Once again, we **expanded the model** in order to answer the question of interest.
- Here, the normalization constant was accessible directly.
  This is very rare and occured because I chose to make my life simple.

Conclusions

The story so far

We now know how to answer key statistical questions of stastical inference:

1. Estimation:
   1. Point estimates:
      - compress the posterior into a scalar: MAP, Mean, Loss function.
   2. Intervals:
      - credible intervals (loss function??).
2. Prediction:
   - **Augment the model.**
   - Obtain a posterior on desired variables.
3. Model selection:
   - **Augment the model.**
   - **Compute the normalization constants in the intermediate posterior.**
   - Obtain a posterior on "Flag" variables.

The story so far

In **all examples so far**, I've made my life **simple**: the posterior was always explicit.

This is **rare**. We'll highlight next week the necessary conditions and explain why this almost never occurs in practice.

Bayesian prediction

Given data $\mathcal{D} = d$ and unobserved variable(s) of interest $Y_{prediction}$:

1. Choose a joint model of data and variable(s) of interest:

$$f\left(\mathcal{D}\ \&\ Y_{prediction}|\boldsymbol{\theta}\right)$$

2. Choose a prior: $f\left(\boldsymbol{\theta}\right)$.
3. Compute the posterior distribution of the random variables $\boldsymbol{\theta}, Y_{prediction}$:

$$f\left(\boldsymbol{\theta}, Y_{prediction}|\mathcal{D} = d\right)$$

4. Marginalize out $\boldsymbol{\theta}$:

$$f\left(Y_{prediction}|\mathcal{D} = d\right) = \int d\boldsymbol{\theta} f\left(\boldsymbol{\theta}, Y_{prediction}|\mathcal{D} = d\right)$$

5. Return Bayesian point or interval estimates of $Y_{prediction}$

Bayesian model selection

Given data $\mathcal{D} = d$ and multiple competing models:

$$f_{M_1}(\mathcal{D}|\boldsymbol{\theta}) \quad f_{M_2}(\mathcal{D}|\boldsymbol{\theta}) \quad f_{M_3}(\mathcal{D}|\boldsymbol{\theta}) \dots$$

(NB: very often, the number or interpretation of the parameters might change drastically from one model to the next):

1. Augment the model with a Flag variable I such that $I = i$ means that model M_i is active.
2. Choose the prior: $\mathbb{P}(I = i)$.
3. For every model, compute the normalization probability of the model:

$$f_{M_i}(\mathcal{D} = d) = \int d\boldsymbol{\theta} f_{M_i}(\mathcal{D} = d|\boldsymbol{\theta})$$

4. The posterior over I is:

$$\mathbb{P}(I = i) \propto \mathbb{P}(I = i) f_{M_i}(\mathcal{D} = d)$$

Methods for prior choice: pseudo-data

The following principles can guide our choice of prior:

1. Vague priors:
   - One weak principle for prior choice is to use priors with very large width.
   - This encodes the common situation of having not much prior information.
2. Pseudo-data interpretation:
   - Priors that are conjugate to a conditional model can be interpreted in terms of adding virtual observations. The value of these virtual observations can be deduced from the parameters of the prior.
   - e.g. Gaussian prior, Beta prior, Student prior.
3. Generative priors:
   - Sampling datasets from the prior distribution $f(\mathcal{D})$ should generate (somewhat) credible artificial datasets.
   - This principle rarely constraints the shape but can be helpful in finding the appropriate scale of the prior.