Lecture 25: Conjugate Priors

Reading: Section 8.3

GU4241/GR5241 Statistical Machine Learning

Linxi Liu
April 20, 2017

Bayesian models

The defining assumption of **Bayesian statistics** is that the distribution P which models the data is a random quantity and itself has a distribution Q. The generative model for data $X_1, X_2, \ldots$ is

$$\begin{aligned} P &\sim Q \\ X_1, X_2, \ldots &\overset{\text{i.i.d.}}{\sim} P \end{aligned}$$

Bayesian models

The defining assumption of **Bayesian statistics** is that the distribution P which models the data is a random quantity and itself has a distribution Q. The generative model for data $X_1, X_2, \ldots$ is

$$\begin{aligned} P &\sim Q \\ X_1, X_2, \ldots &\overset{\text{i.i.d.}}{\sim} P \end{aligned}$$

The rational behind the approach is:

- In any statistical approach (Bayesian or frequentist), the distribution P is unknown.

Bayesian models

The defining assumption of **Bayesian statistics** is that the distribution P which models the data is a random quantity and itself has a distribution Q. The generative model for data $X_1, X_2, \ldots$ is

$$\begin{aligned} P &\sim Q \\ X_1, X_2, \ldots &\overset{\text{i.i.d.}}{\sim} P \end{aligned}$$

The rational behind the approach is:

- In any statistical approach (Bayesian or frequentist), the distribution P is unknown.
- Bayesian statistics argues that any form of uncertainty should be expressed by probability distributions.

Bayesian models

The defining assumption of **Bayesian statistics** is that the distribution P which models the data is a random quantity and itself has a distribution Q. The generative model for data $X_1, X_2, \ldots$ is

$$\begin{aligned} P &\sim Q \\ X_1, X_2, \ldots &\overset{\text{i.i.d.}}{\sim} P \end{aligned}$$

The rational behind the approach is:

- In any statistical approach (Bayesian or frequentist), the distribution P is unknown.
- Bayesian statistics argues that any form of uncertainty should be expressed by probability distributions.
- We can think of the randomness in Q as a model of the statistician's lack of knowledge regarding P.

Prior and posterior

The distribution Q of P is called the **a priori distribution** (or the **prior** for short). We use q to denote its density if it exists.

Prior and posterior

The distribution Q of P is called the **a priori distribution** (or the **prior** for short). We use q to denote its density if it exists.

Our objective is to determine the conditional probability of P given observed data

$$\mathsf{Pr}(P|x_1, \ldots, x_n).$$

Prior and posterior

The distribution Q of P is called the **a priori distribution** (or the **prior** for short). We use q to denote its density if it exists.

Our objective is to determine the conditional probability of P given observed data

$$\Pr(P|x_1, \ldots, x_n).$$

The distribution is called the **a posteriori distribution** or **posterior**.

Parametric Case

We can impose the modeling assumption that P is an element of a parametric model, e.g. that the density p of P is in a family $\mathcal{P} = \{p(x|\theta)|\theta \in \mathcal{T}\}$. If so, the prior and posterior can be expressed as distributions on $\mathcal{T}$. We write

$$q(\theta) \qquad \text{and} \qquad \text{Pr}(\theta|x_1, \dots, x_n)$$

for the prior and posterior density, respectively.

Remark

The posterior $\text{Pr}[P|x_1, \dots, x_n]$ is an abstract object, which can be rigorously defined using the tools of probability theory, but is in general (even theoretically) impossible to compute. However: In the parametric case, the posterior can be obtained using the Bayes equation.

Bayes' Theorem

Parametric modeling assumption

Suppose $\mathcal{P} = \{p(x|\theta)|\theta \in \mathcal{T}\}$ is a model and q a prior distribution on $\mathcal{T}$. Our sampling model then has the form:

$$\begin{aligned} \theta &\sim q \\ X_1, X_2, \ldots &\overset{\text{i.i.d.}}{\sim} p(\,.\,|\theta) \end{aligned}$$

Note that the data is *conditionally i.i.d.* given $\Theta = \theta$.

Bayes' Theorem

Parametric modeling assumption

Suppose $\mathcal{P} = \{p(x|\theta)|\theta \in \mathcal{T}\}$ is a model and q a prior distribution on $\mathcal{T}$. Our sampling model then has the form:

$$\begin{aligned} \theta &\sim q \\ X_1, X_2, \ldots &\overset{\text{i.i.d.}}{\sim} p(\,.\,|\theta) \end{aligned}$$

Note that the data is *conditionally i.i.d.* given $\Theta = \theta$.

Given data $X_1, \ldots, X_n$, we can compute the posterior by

$$\Pr(\theta|x_1, \ldots, x_n) = \frac{(\prod_{i=1}^{n} p(x_i|\theta))q(\theta)}{p(x_1, \ldots, x_n)} = \frac{(\prod_{i=1}^{n} p(x_i|\theta))q(\theta)}{\int (\prod_{i=1}^{n} p(x_i|\theta))\, q(\theta)}.$$

this is very difficult to calculate

The individual terms have names: We only know the posterior distribution

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

Example: unknown Gaussian mean

Model

We assume that the data is generated from a Gaussian with fixed variance σ^2. The mean μ is unknown. The model likelihood is $p(x|\mu, \sigma) = g(x|\mu, \sigma)$ (where g is the Gaussian density on the line).

Example: unknown Gaussian mean

Model

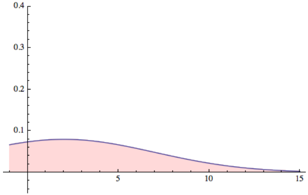
We assume that the data is generated from a Gaussian with fixed variance σ^2. The mean μ is unknown. The model likelihood is $p(x|\mu, \sigma) = g(x|\mu, \sigma)$ (where g is the Gaussian density on the line).

Bayesian model

We choose a Gaussian prior on μ,

$$q(\mu) := g(\mu|\mu_0, \sigma_0) \ .$$

In the figure, $\mu_0 = 2$ and $\sigma_0 = 5$. Hence, we assume that $\mu_0 = 2$ is the most probable value of μ, and that $\mu \in [-3, 7]$ with a probability ~ 0.68.



Example: Unknown Gaussian mean

Application of Bayes' formula to the Gaussian-Gaussian model shows the posterior distribution is

$$\Pr(\mu | x_{1:n}) = g(\mu | \mu_n, \sigma_n),$$

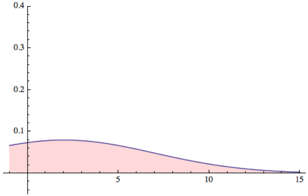
where $\mu_n := \frac{\sigma^2 \mu_0 + \sigma_0^2 \sum_{i=1}^n x_i}{\sigma^2 + n\sigma_0^2}$ and $\sigma_n^2 := \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$.

Example: Unknown Gaussian mean

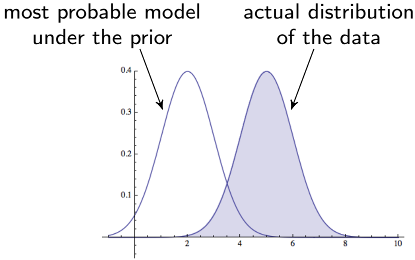
Application of Bayes' formula to the Gaussian-Gaussian model shows the posterior distribution is

$$\Pr(\mu | x_{1:n}) = g(\mu | \mu_n, \sigma_n),$$

where $\mu_n := \frac{\sigma^2 \mu_0 + \sigma_0^2 \sum_{i=1}^n x_i}{\sigma^2 + n\sigma_0^2}$ and $\sigma_n^2 := \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$.



Prior



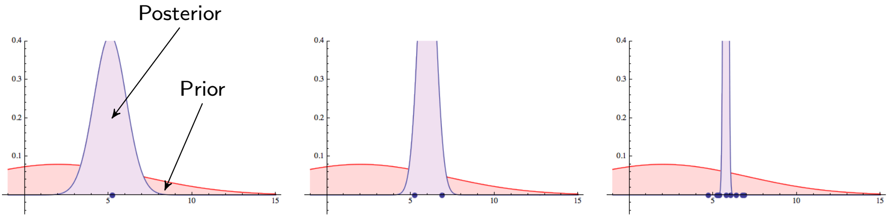
Sampling distribution

Example: Unknown Gaussian mean

Application of Bayes' formula to the Gaussian-Gaussian model shows the posterior distribution is

$$\Pr(\mu|x_{1:n}) = g(\mu|\mu_n, \sigma_n),$$

where $\mu_n := \frac{\sigma^2\mu_0 + \sigma_0^2 \sum_{i=1}^n x_i}{\sigma^2 + n\sigma_0^2}$ and $\sigma_n^2 := \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}$.



$n = 1$ $n = 2$ $n = 10$

Exponential Family Distributions

Definition

We consider a model $\mathcal{P}$ for data in a sample space $\mathbf{X}$ with parameter space $\mathcal{T} \subset \mathbb{R}^m$. Each distribution in $\mathcal{P}$ has density $p(x|\theta)$ for some $\theta \in \mathcal{T}$.

The model is called an **exponential family model** (EFM) if p can be written as

$$p(x|\theta) = \frac{h(x)}{Z(\theta)} e^{\langle S(x), \theta \rangle}$$

where:

- S is a function $S : \mathbf{X} \to \mathbb{R}^m$. This function is called the **sufficient statistic** of $\mathcal{P}$.
- h is a function $h : \mathbf{X} \to \mathbb{R}_+$.
- Z is a function $Z : \mathcal{T} \to \mathbb{R}_+$, called the **partition function**.

Exponential Family Distributions

Exponential families are important because:

1. The special form of p gives them many nice properties.
2. Most important parametric models (e.g. Gaussians) are EFMs.
3. Many algorithms and methods can be formulated generically for all EFMs.

Alternative Form

The choice of p looks perhaps less arbitrary if we write

$$p(x|\theta) = \exp\Big(\langle S(x), \theta \rangle - \phi(x) - \psi(\theta)\Big)$$

which is obtained by defining

$$\phi(x) := -\log(h(x)) \qquad \text{and} \qquad \psi(\theta) := \log(Z(\theta))$$

A first interpretation

Exponential family models are models in which:

- The data and the parameter interact only through the linear term $\langle S(x), \theta \rangle$ in the exponent.

Alternative Form

The choice of p looks perhaps less arbitrary if we write

$$p(x|\theta) = \exp\Big(\langle S(x), \theta\rangle - \phi(x) - \psi(\theta)\Big)$$

which is obtained by defining

$$\phi(x) := -\log(h(x)) \qquad \text{and} \qquad \psi(\theta) := \log(Z(\theta))$$

A first interpretation

Exponential family models are models in which:

- The data and the parameter interact only through the linear term $\langle S(x), \theta\rangle$ in the exponent.
- The logarithm of p can be non-linear in both $S(x)$ and θ, but there is no *joint* nonlinear function of $(S(x), \theta)$.

The Partition Function

Normalization constraint

Since p is a probability density, we know

$$\int_{\mathbf{X}} \frac{h(x)}{Z(\theta)} e^{\langle S(x),\theta\rangle} dx = 1 \ .$$

Partition function

The only term we can pull out of the integral is the partition function $Z(\theta)$, hence

$$Z(\theta) = \int_{\mathbf{X}} h(x) e^{\langle S(x),\theta\rangle} dx$$

Note: This implies that an exponential family is completely determined by choice of the spaces $\mathbf{X}$ and $\mathcal{T}$ and of the functions S and h.

Example: Gaussian

In 1 dimension

We can rewrite the exponent of the Gaussian as

$$\begin{aligned}\frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{1}{2}\frac{x^2}{\sigma^2}+\frac{2x\mu}{2\sigma^2}\Big)\exp\Big(-\frac{1}{2}\frac{\mu^2}{\sigma^2}\Big)\\ &= \underbrace{c(\mu,\sigma)}_{\text{some function of } \mu \text{ and } \sigma}\exp\Big(x^2\cdot\frac{-1}{2\sigma^2}+x\cdot\frac{\mu}{\sigma^2}\Big)\end{aligned}$$

This shows the Gaussian is an exponential family, since we can choose:

$S(x) := \big(x^2, x\big)$ and $\theta := \big(\frac{-1}{2\sigma^2}, \frac{\mu}{\sigma^2}\big)$ and $h(x) = 1$ and $Z(\theta) = c(\mu,\sigma)^{-1}$.

In d dimensions

$$S(\mathbf{x}) = \big(\mathbf{x}\mathbf{x}^t, \mathbf{x}\big) \qquad \text{and} \qquad \theta := \big(-\tfrac{1}{2}\Sigma^{-1}, \Sigma^{-1}\mu\big)$$

Back to Bayesian Models: Parametric Prior Families

Families of priors

The prior has to be expressed by a specific distribution. In parametric Bayesian models, we typically choose q as an element of a standard parametric family (e.g. the Gaussian in the previous example).

Hyperparameters

Text

If we choose q as an element of a parametric family

$$\mathcal{Q} = \{q(\theta|\phi)|\phi \in \mathcal{H}\}$$

Φ is the parameter of the prior

on $\mathcal{T}$, selecting the prior comes down to choosing ϕ. Hence, ϕ becomes a tuning parameter of the model.

$q(\theta|\lambda,y) = \exp\langle\theta,y\rangle - \lambda Z(\theta)$

posterior: $\exp(\langle\Theta,S(x) + y\rangle - (\lambda+n)\log Z(\theta)$

Parameter of the prior familiy are called **hyperparameters** of the Bayesian model.

Natural Conjugate Priors

Exponential family likelihood

We now assume the parametric model $\mathcal{P} = \{p(x|\theta)|\theta \in \mathcal{T}\}$ is an exponential family model, i.e.

$$p(x|\theta) = \frac{h(x)}{Z(\theta)} e^{\langle S(x)|\theta \rangle} \ .$$

Natural conjugate prior

We define a prior distribution using the density

$$q(\theta|\lambda, y) = \frac{1}{K(\lambda, y)} \exp\Big(\langle \theta | y \rangle - \lambda \cdot \log Z(\theta) \Big)$$

- Hyperparameters: $\lambda \in \mathbb{R}_+$ and $y \in \mathcal{T}$.
- Note that the choice of $\mathcal{P}$ enters through Z.
- K is a normalization function.

Clearly, this is itself an exponential family (on $\mathcal{T}$), with $h \equiv Z^{-\lambda}$ and $Z \equiv K$.

Ugly Computation

Substitution into Bayes' equation gives

$$\mathsf{Pr}(\theta|x_1,\ldots,x_n) = \frac{\prod_{i=1}^n p(x_i|\theta)}{p(x_1,\ldots,x_n)} \cdot q(\theta)$$
$$= \frac{\frac{\prod_{i=1}^n h(x_i)}{Z(\theta)^n} \exp\left\langle \sum_i S(x_i)|\theta \right\rangle}{p(x_1,\ldots,x_n)} \cdot \frac{\exp\big(\langle\theta|y\rangle - \lambda \log Z(\theta)\big)}{K(\lambda,y)}$$

If we neglect all terms which do not depend on θ, we have

$$\mathsf{Pr}(\theta|x_1,\ldots,x_n) \propto \; = \frac{\exp\left\langle \sum_i S(x_i)|\theta \right\rangle}{Z(\theta)^n} \exp\big(\langle\theta|y\rangle - \lambda \log Z(\theta)\big) = \frac{\exp\Big(\left\langle y + \sum_i S(x_i)|\theta \right\rangle\Big)}{Z(\theta)^{\lambda+n}}$$

Up to normalization, this is precisely the form of an element of $\mathcal{Q}$:

$$\ldots = \exp\Bigg(\Bigg\langle y + \sum_i S(x_i)|\theta \Bigg\rangle - (\lambda+n)\log Z(\theta)\Bigg) \propto q(\theta|\lambda+n, y + \sum_{i=1}^n S(x_i))$$

Posteriors of Conjugate Priors

Conclusion

If $\mathcal{P}$ is an exponential family model with sufficient statistic S, and if $q(\theta|\lambda, y)$ is a natural conjugate prior for $\mathcal{P}$, the posterior under observations $x_1, \ldots, x_n$ is

$$\text{Pr}(\theta|x_1, \ldots, x_n) = q(\theta|\lambda + n, y + \sum_{i=1}^{n} S(x_i))$$

Remark

The form of the posterior above means that we can *compute the posterior by updating the hyperparameters*. This property motivates the next definition.

Definition

Assume that $\mathcal{P}$ is a parametric family and $\mathcal{Q}$ a family of priors. Suppose, for each sample size $n \in \mathbb{N}$, there is a function $T_n : \mathbf{X}^n \times \mathcal{H} \to \mathcal{H}$ such that

$$\text{Pr}(\theta|x_1, \ldots, x_n) = q(\theta|\hat{\phi}) \qquad \text{with} \qquad \hat{\phi} := T_n(x_1, \ldots, x_n, \phi)\ .$$

Then $\mathcal{P}$ and $\mathcal{Q}$ are called **conjugate**.

Conjugate Priors

Closure under sampling

If the posterior is an element of the prior family, i.e. if

$$\Pr(\theta|x_1, \ldots, x_n) = q(\theta|\tilde{\phi})$$

for *some* $\tilde{\phi}$, the model is called **closed under sampling**. Clearly, every conjugate model is closed under sampling.

Remark

Closure under sampling is a weaker property than conjugacy; for example, any Bayesian model with

$$\mathcal{Q} = \{ \text{ all probability distributions on } \mathcal{T} \}$$

is trivially closed under sampling, but not conjugate.
Warning: Many Bayesian texts use conjugacy and closure under sampling equivalently.

Which models are conjugate?

It can be shown that, up a few "borderline" cases, the only paramteric models which admit conjugate priors are exponential family models.