Applied Probabilistic Programming & Bayesian Machine Learning

§2-3: Robust Models

MIT IAP 2017 Max W. Shen

1 Robust Models

One property of Bayesian models is that "importance weighting" of some samples is not easily done, since Bayesian models assume random variables as equally important random samples from a distribution. Here, we show how to create models that are robust to outliers.

Exercise 1

Simulate data with outliers. Use a model that assumes a normal distribution over data. Compare inference results on data with the outliers, and data after prefiltering and removing the outliers.

1.1 Normal distribution and *t*-distribution

The normal distribution is notoriously sensitive to outliers in data. In hierarchical models with interacting components, outliers can disrupt inference on parameters throughout the model.

One solution is to replace the normal distribution with more *long-tailed* distributions. Placing more probability mass in the tails enables the model to be more forgiving of outliers. An advantage to this approach is preserving fully Bayesian analysis.

In essence, we replace the original distribution with a distribution that has greater variance.

The t-distribution shares a location parameter μ and non-negative scale parameter σ in common with the normal distribution, but has an additional parameter $\nu > 0$ for degrees of freedom. When $\nu = \infty$, the t-distribution is equivalent to the normal distribution, and when $\nu = 1$, the t-distribution is equivalent to a Cauchy distribution.

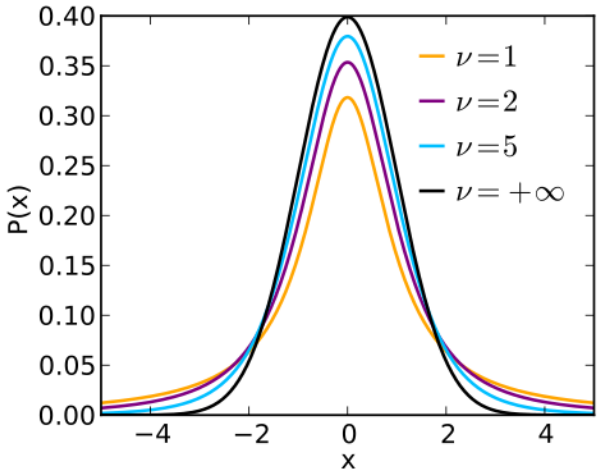

Figure 1: The t-distribution at varying ν

Decreasing ν increases the variance of the t-distribution and flattens it out.

Distribution	Notation	Parameters	Density function	Mean, variance, and mode
t	$\theta \sim t_\nu(\mu, \sigma^2)$ $p(\theta) = t_\nu(\theta \mid \mu, \sigma^2)$ t_ν is short for $t_\nu(0, 1)$	degrees of freedom $\nu > 0$ location μ scale $\sigma > 0$	$p(\theta) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma}(1 + \frac{1}{\nu}(\frac{\theta-\mu}{\sigma})^2)^{-(\nu+1)/2}$	$\mathrm{E}(\theta) = \mu$, for $\nu > 1$ $\mathrm{var}(\theta) = \frac{\nu}{\nu-2}\sigma^2$, for $\nu > 2$ $\mathrm{mode}(\theta) = \mu$
Normal	$\theta \sim \mathrm{N}(\mu, \sigma^2)$ $p(\theta) = \mathrm{N}(\theta \mid \mu, \sigma^2)$	location μ scale $\sigma > 0$	$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\theta - \mu)^2\right)$	$\mathrm{E}(\theta) = \mu$ $\mathrm{var}(\theta) = \sigma^2$ $\mathrm{mode}(\theta) = \mu$

Figure 2: t-distribution and Normal distribution

From figure 2, the variance of the t-distribution is defined for $\nu > 2$, so in practice we suggest using $\nu > 2$ for robust analysis. Notably, the Cauchy distribution (corresponding to $\nu = 1$) does not have finite variance.

As a brief aside, the popular data visualization technique t-SNE (where t refers to the t-distribution) takes advantage of the longer tails of the t-distribution to address geometric constraints that occur when projecting high-dimensional data points into two or threee dimensions. Informally speaking, by relating probability to distance, t-SNE allows nearby points in high-dimensional space to be represented at greater distances in 2D or 3D.

1.1.1 Mixture Interpretation

Models 1 and 2 are equivalent.

$$y_i \sim t_\nu(\mu, \sigma^2) \tag{1}$$

$$\begin{aligned} y_i &\sim N(\mu, V_i) \\ V_i &\sim Inv - \chi^2(\nu, \sigma^2) \end{aligned} \tag{2}$$

Model 2 can be interpreted as saying the population of data y is distributed according to a mixture of normal distributions, each with varying scale (variance, σ^2). This is known as a scale mixture of normals. The variances are drawn from a common Inverse-χ^2 distribution which uses the same ν and σ^2 parameters from model 1.

Since each datapoint y_i gets its own variance V_i, the model is granted greater flexibility to model outliers.

Exercise 2

Using the model you developed with outliers, replace the normal distribution with a t-distribution as in model 1 and a scale mixture of normals as in 2. Compare your modeling results with the normal distribution with and without the outliers.

1.2 Poisson distribution and Negative binomial distribution

Distribution	Notation	Parameters	Density function	Mean, variance, and mode
Poisson	$\theta \sim \text{Poisson}(\lambda)$ $p(\theta) = \text{Poisson}(\theta\|\lambda)$	'rate' $\lambda > 0$	$p(\theta) = \frac{1}{\theta!}\lambda^\theta \exp(-\lambda)$ $\theta = 0, 1, 2, \ldots$	$\text{E}(\theta) = \lambda$, $\text{var}(\theta) = \lambda$ $\text{mode}(\theta) = \lfloor\lambda\rfloor$
Distribution	**Notation**	**Parameters**	**Density function**	**Mean, variance, and mode**
Negative binomial	$\theta \sim \text{Neg-bin}(\alpha, \beta)$ $p(\theta) = \text{Neg-bin}(\theta\|\alpha, \beta)$	shape $\alpha > 0$ inverse scale $\beta > 0$	$p(\theta) = \binom{\theta+\alpha-1}{\alpha-1}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^\theta$ $\theta = 0, 1, 2, \ldots$	$\text{E}(\theta) = \frac{\alpha}{\beta}$ $\text{var}(\theta) = \frac{\alpha}{\beta^2}(\beta+1)$

Figure 3: The Poisson and Negative-Binomial Distributions

The Poisson distribution has just one parameter, λ, which determines both its mean and variance. In practice, real data may best be modeled with a larger variance than the Poisson distribution can support given a particular mean.

The negative binomial distribution, given a mean $\lambda = \frac{\alpha}{\beta}$, always has a larger variance than the Poisson distribution. The negative binomial distribution converges to the Poisson distribution with fixed $\frac{\alpha}{\beta}$ as $\beta \to \infty$.

1.2.1 Mixture Interpretation

Models 3 and 4 are equivalent.

$$y_i \sim NegBin(\alpha, \beta) \tag{3}$$

$$\begin{aligned} y_i &\sim Poisson(\lambda_i) \\ \lambda_i &\sim Gamma(\alpha, \beta) \end{aligned} \tag{4}$$

Distribution	Notation	Parameters	Density function	Mean, variance, and mode
Binomial	$\theta \sim \text{Bin}(n, p)$ $p(\theta) = \text{Bin}(\theta \mid n, p)$	'sample size' n (positive integer) 'probability' $p \in [0, 1]$	$p(\theta) = \binom{n}{\theta} p^{\theta}(1-p)^{n-\theta}$ $\theta = 0, 1, 2, \ldots, n$	$\text{E}(\theta) = np$ $\text{var}(\theta) = np(1-p)$ $\text{mode}(\theta) = \lfloor (n+1)p \rfloor$
Distribution	Notation	Parameters	Density function	Mean, variance, and mode
Beta-binomial	$\theta \sim \text{Beta-bin}(n, \alpha, \beta)$ $p(\theta) = \text{Beta-bin}(\theta \mid n, \alpha, \beta)$	'sample size' n (positive integer) 'prior sample sizes' $\alpha > 0, \beta > 0$	$p(\theta) = \frac{\Gamma(n+1)}{\Gamma(\theta+1)\Gamma(n-\theta+1)} \frac{\Gamma(\alpha+\theta)\Gamma(n+\beta-\theta)}{\Gamma(\alpha+\beta+n)}$ $\times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}, \quad \theta = 0, 1, 2, \ldots, n$	$\text{E}(\theta) = n\frac{\alpha}{\alpha+\beta}$ $\text{var}(\theta) = n\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Figure 4: The Binomial and Beta-Binomial Distributions

1.3 Binomial distribution and the Beta-binomial distribution

Exercise 3

Show that the beta-binomial's variance is greater by $\frac{\alpha+\beta+n}{\alpha+\beta+1}$ than the binomial's variance, when $p = \frac{\alpha}{\alpha+\beta}$.

1.3.1 Mixture Interpretation

Let there be n datapoints, indexed by i. That is, we have $y_1, ..., y_i, ..., y_n$. Then, models 5 and 6 are equivalent.

$$y_i \sim Beta - bin(m, \alpha, \beta) \tag{5}$$

$$\begin{aligned} y_i &\sim Binomial(p_i, m) \\ p_i &\sim Beta(\alpha, \beta) \end{aligned} \tag{6}$$

Exercise 4

Write the non-robust / standard version of model 5. (Use statistical notation on paper, no need to write code).

2 Bibliography and Additional Resources

Chapter 17 of Bayesian Data Analysis 3 discusses robust Bayesian modeling.

The Stan reference is available here: http://mc-stan.org/documentation/