

Байесовское мультимоделирование: байесовский вывод

Московский Физико-Технический Институт

2021

Задача о монетке

Человек подбрасывает монетку 3 раза. Все 3 раза выпадает решка. Оценить вероятность выпадения решки на монетке.

Наивный подход

$$\mathbf{X} = [1, 1, 1];$$

$$x \sim \text{Bin}(w);$$

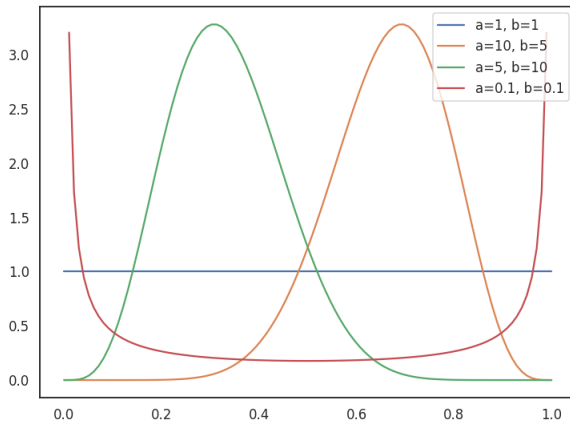
$$\hat{w} = \arg \max_p L(\mathbf{X}, w);$$

$$\rightarrow \hat{w} = 1.$$

Проблема: трех событий может быть недостаточно для оценки распределения орлов и решек.

Бета-распределение: напоминание

- соответствует *априорным* ожиданиям о распределении Бернулли
- интерпретация: “эффективное количество наблюдений $w = 1, w = 0$ ”
- при $n \rightarrow \infty$ сходится к δ -распределению в точке ОМП распределения Бернулли.



Байесовский подход

Введем бета-распределение в качестве *априорного* предположения о распределении нашего параметра. Из общих соображений распределение должно быть симметрично (если у нас нет дополнительной информации):

$$p(w) \sim B(\alpha, \beta).$$

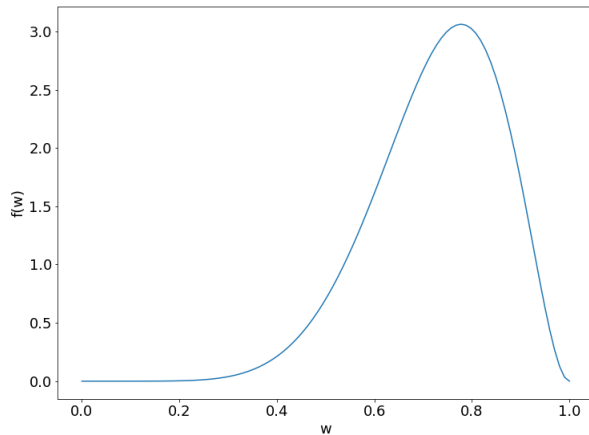
Найдем *апостериорное* распределение параметра w распределения Бернулли по формуле Байеса:

$$p(w|\mathbf{x}) = \frac{p(\mathbf{X}|w)p(w)}{p(\mathbf{X})} \propto p(\mathbf{X}|w)p(w);$$

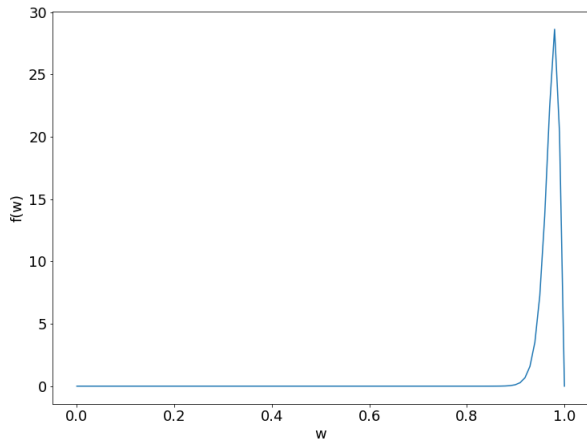
$$\log p(w|\mathbf{x}) = \log p(\mathbf{X}|w) + \log p(w) + \text{Const.}$$

Вывод: грубая интерпретация априорного распределения — *регуляризатор*.

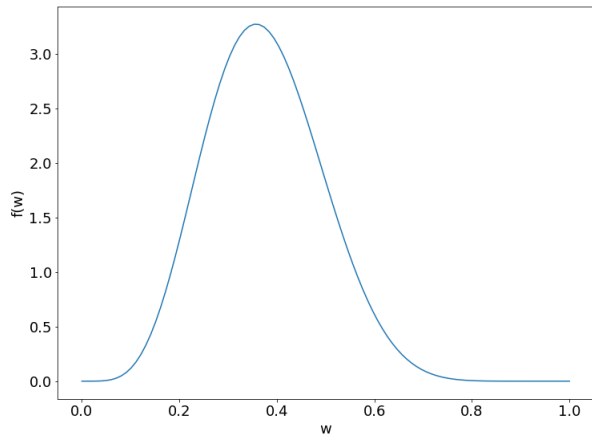
Апостериорное распределение, $\alpha = 3, \beta = 3$



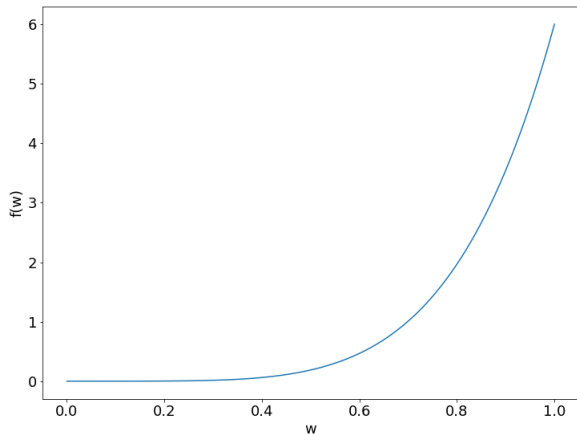
Апостериорное распределение, выборка из 100 элементов



Апостериорное распределение, $\alpha = 1, \beta = 10$



Апостериорное распределение, $\alpha = 1, \beta = 1$



Байесовский вывод: первый уровень

Заданы:

- правдоподобие $p(\mathbf{X}|\mathbf{w})$ выборки \mathbf{X} при условии параметра \mathbf{w} ;
- априорное распределение $p(\mathbf{w}|\mathbf{h})$
- параметры априорного распределения \mathbf{h} (В примере с монеткой: $\mathbf{h} = [\alpha, \beta]$;))

Тогда апостериорное распределение параметров \mathbf{w} при условии выборки \mathbf{X} :

$$p(\mathbf{w}|\mathbf{x}, \mathbf{h}) = \frac{p(\mathbf{X}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathbf{X}|\mathbf{h})} \propto p(\mathbf{X}|\mathbf{w})p(\mathbf{w}|\mathbf{h}).$$

Точечная оценка параметров находится как максимум апостериорной вероятности (MAP):

$$\hat{\mathbf{w}} = \arg \max p(\mathbf{X}|\mathbf{w})p(\mathbf{w}|\mathbf{h}).$$

MAP-оценка схожа с оценкой методом максимального правдоподобия, если

- Мощность выборка велика
- Априорное распределение — равномерное на очень большой области

Почему для монетки подошло бета-распределение?

$$\begin{aligned} p(w|\mathbf{x}, \alpha, \beta) &\propto p(\mathbf{X}|w)p(w|\alpha, \beta) \propto \\ &\propto w^{\sum x}(1-w)^{m-\sum x} \times w^{\alpha-1}(1-w)^{\beta-1} = \\ &= w^{\alpha-1+\sum x}(1-w)^{m+\beta-\sum x-1} \sim B(\alpha + \sum x, \beta + m - \sum x). \end{aligned}$$

Семейство распределений называется сопряженным к распределению правдоподобия, если апостериорное распределение принадлежит этому же семейству.

Априорные распределения

- Для дискретных величин (отклики, дискретные параметры)

- ▶ Распределение Бернулли
- ▶ Категориальное распределение

Гиперпараметры:

- ▶ $w \sim \text{Bin}(w)$: $w \sim B(\alpha, \beta)$: сопряженное распределение
- ▶ $w \sim \text{Cat}(w)$: $w \sim \text{Dir}(\alpha)$: сопряженное распределение

- Для вещественнозначных величин

- ▶ \mathcal{N}
- ▶ Laplace
- ▶ \mathcal{C}

Гиперпараметры:

- ▶ Дисперсия, $w \sim \mathcal{N}(\mu, \sigma^2)$, $\sigma^2 \in \Gamma$: сопряженное нормальному распределению
- ▶ Матожидание, $\mu \in \mathcal{N}$: сопряженное нормальному распределению

Informative prior vs Uninformative prior

- Informative prior: соответствует экспертным знаниям о наблюдаемой переменной
 - ▶ Пример: температура воздуха: нормальная величина с известным средним и дисперсией, соответствующими прошлым наблюдениям.
 - ▶ Соответствие апостериорного распределения априорному назовем интерпретируемостью модели.
 - ▶ Ошибка в указании информативного априорного распределения может значительно снизить итоговое качество модели.
- Uninformative prior: соответствует базовым предположениям о распределении переменной
 - ▶ Пример: температура воздуха: равномерное распределение (improper).
- Weakly-informative prior: где-то по середине
 - ▶ Пример: температура воздуха: равномерное распределение от -50 до +50.

Вопросы:

- $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$ — какой тип априорного распределения?
- Как интерпретировать соответствие параметра его априорному распределению?

Априорное распределение Джеффриса

Неинформативное распределение следующего вида:

$$p(\mathbf{w}) \propto \sqrt{\det I(\mathbf{w})} = \sqrt{\det \left(-\frac{\partial^2}{\partial w^2} \log L(w) \right)}.$$

- Распределение инвариантно относительно замены переменных:

$$p(g(\mathbf{w})) = p(\mathbf{w}) \left| \frac{dg}{d\mathbf{w}} \right| \rightarrow$$

$$p(g(\mathbf{w})) \propto \sqrt{\det I(g(\mathbf{w}))}.$$

- Интерпретация: величина, обратно пропорциональная информации, получаемой моделью из выборки
- Примеры распределений:
 - ▶ $y \in \text{Bin}(w) : p(w) \propto \frac{1}{\sqrt{p(1-p)}}$ — Beta-распределение с параметрами (0.5, 0.5).
 - ▶ $w \in \mathcal{N}(\mu, \sigma) : p(\mu) \propto \text{Const.}$
 - ▶ $w \in \mathcal{N}(\mu, \sigma) : p(\sigma) \propto \frac{1}{|\sigma|}.$

Выбор модели: связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

$$\mathbf{w} = \arg \max \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathcal{D}|\mathbf{h})},$$

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели.

Обоснованность модели (“Evidence”):

$$p(\mathcal{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathcal{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$

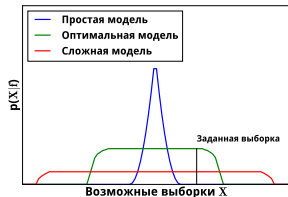
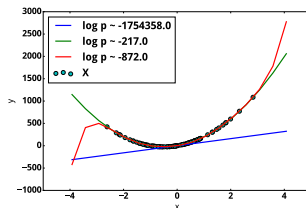


Схема выбора модели



Пример: полиномы

Пример: линейная регрессия

Линейный случай с m объектами и n признаками: $\mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w}$;
 $\mathbf{y} \sim \mathcal{N}(\mathbf{f}(\mathbf{X}, \mathbf{w}), \beta^{-1})$, $\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1})$.

Запишем интеграл:

$$\begin{aligned} p(\mathcal{D}|\mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \beta) &= \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp(-0.5\beta(\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f})) \exp(-0.5\mathbf{w}^T \mathbf{A} \mathbf{w}) d\mathbf{w} = \\ &= \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp(-S(\mathbf{w})) d\mathbf{w} \end{aligned}$$

Для линейного случая интеграл вычисляется аналитически:

$$\int_{\mathbf{w}} \exp(-S(\mathbf{w})) d\mathbf{w} = (2\pi)^{\frac{n}{2}} S(\hat{\mathbf{w}}) |\mathbf{H}^{-1}|^{0.5},$$

где

$$\begin{aligned} \mathbf{H} &= \mathbf{A} + \beta \mathbf{X}^T \mathbf{X}, \\ \hat{\mathbf{w}} &= \beta \mathbf{H}^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

Вывод: для линейных моделей Evidence считается аналитически.

Пример: аппроксимация Лапласа

Нелинейный случай с m объектами и n признаками: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}(\mathbf{X}, \mathbf{w}), \beta^{-1})$, $\mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1})$.
Запишем интеграл:

$$p(\mathcal{D}|\mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \beta) = \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp(-S(\mathbf{w})) d\mathbf{w}.$$

Разложим S в ряд Тейлора:

$$S(\mathbf{w}) \approx S(\hat{\mathbf{w}}) + \frac{1}{2} \Delta \mathbf{w}^T \mathbf{H} \Delta \mathbf{w}$$

Интеграл приводится к виду:

$$\frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} S(\hat{\mathbf{w}}) \int_{\mathbf{w}} \exp\left(-\frac{1}{2} \Delta \mathbf{w}^T \mathbf{H} \Delta \mathbf{w}\right) d\mathbf{w}$$

Выражение под интегралом соответствует плотности ненормированного нормального распределения.

Вывод: для нелинейных моделей можно использовать аппроксимацию Лапласа для получения оценок Evidence.

Литература и прочие ресурсы

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Kuznetsov M., Tokmakova A., Strijov V. Analytic and stochastic methods of structure parameter estimation //Informatica. – 2016. – Т. 27. – №. 3. – С. 607-624.
- Пример с монеткой:
<https://towardsdatascience.com/visualizing-beta-distribution-7391c18031f1>
- Немного про распределение Джеффриса:
<https://medium.datadriveninvestor.com/firths-logistic-regression-classification-with-datasets-that-are-small-imbalanced-or-separated-49d7782a13f1>