Байесовское мультимоделирование: байесовский вывод

Московский Физико-Технический Институт

2021

Задача о монетке

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

Наивный подход

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

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

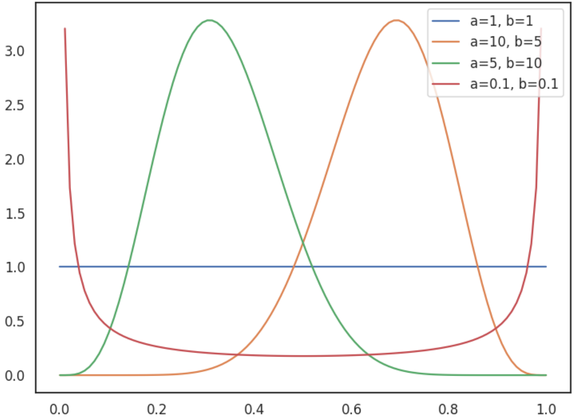
$$\hat{w} = \arg\max_{p} L(\boldsymbol{X}, w);$$

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

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 $n \to \infty$ сходится к δ-распределению в точке ОМП распределения Бернулли.



Байесовский подход

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

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

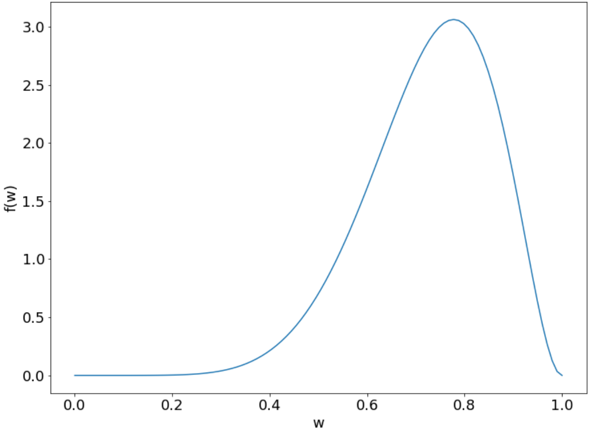
Найдем *апостериорное* распрделение параметра w распределения Бернулли по формуле Байеса:

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

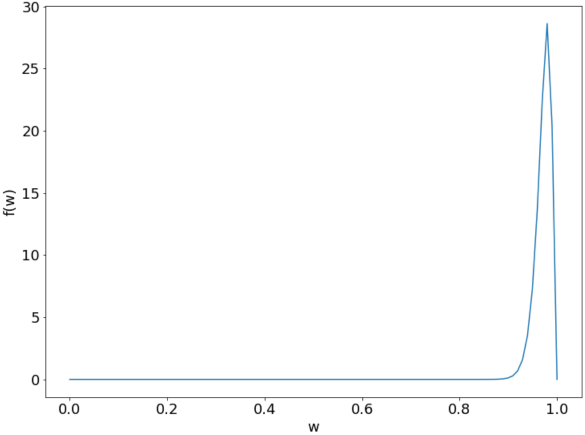
$$\log p(w|\boldsymbol{x}) = \log p(\boldsymbol{X}|w) + \log p(w) + \text{Const}.$$

Вывод: грубая интерпретация априорного распределения — регуляризатор.

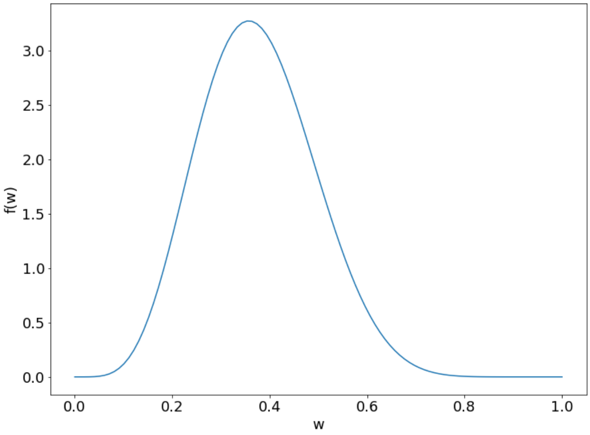
Апостериорное распределение, $\alpha = 3, \beta = 3$



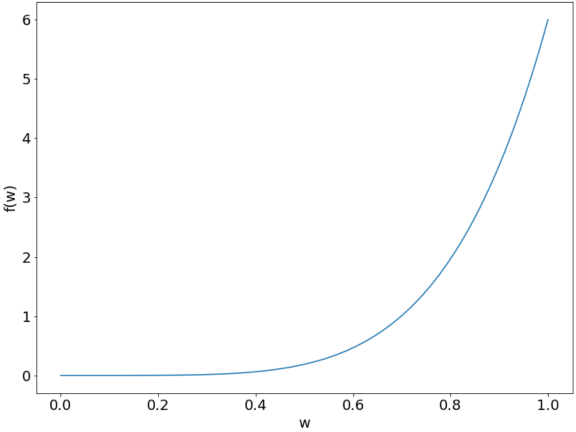
Апостериорное распределение, выборка из 100 элементов



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



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



Байесовский вывод: первый уровень

Заданы:

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

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

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

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

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

MAP-оценка схожа с оценкой методом максимального правдоподобия, если

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

Почему для монетки подошло бета-распределение?

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

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

Априорные распределения

- Для дикретных величин (отклики, дисретные параметры)
  - Распределенеие Бернулли
  - Категориальное распределение

  Гиперпараметры:
  - $w \sim \text{Bin}(w)$: $w \sim B(\alpha, \beta)$: сопряженное распределение
  - $w \sim \text{Cat}(w)$: $w \sim \text{Dir}(\boldsymbol{\alpha})$: сопряженное распределение
- Для вещественнозначных величин
  - $\mathcal{N}$
  - Laplace
  - $\mathcal{C}$

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

- $\boldsymbol{w} \sim \mathcal{N}(0, \boldsymbol{A}^{-1})$ — 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:

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

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

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

Выбор модели: связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

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

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели. Обоснованность модели (“Evidence”):

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

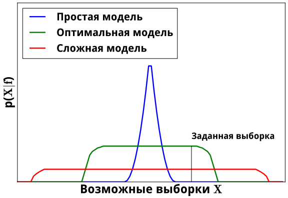
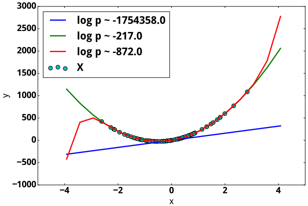


Схема выбора модели



Пример: полиномы

Пример: линейная регрессия

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

$$p(\mathfrak{D}|\boldsymbol{h}) = p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{A}, \beta) = \frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\boldsymbol{w}} \exp\left(-0.5\beta(\boldsymbol{y} - \boldsymbol{f})^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{f})\right) \exp\left(-0.5\beta \boldsymbol{w}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{w}\right) d\boldsymbol{w} =$$

$$= \frac{\sqrt{\beta \cdot |\boldsymbol{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\boldsymbol{w}} \exp(-S(\boldsymbol{w}))d\boldsymbol{w}$$

Для линейного случая интеграл вычисляется аналитически:

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

где

$$\boldsymbol{H} = \boldsymbol{A} + \beta \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X},$$

$$\hat{\boldsymbol{w}} = \beta \boldsymbol{H}^{-1}\boldsymbol{X}^{\mathsf{T}}\boldsymbol{y}$$

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:

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

Разложим S в ряд Тейлора:

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

Интеграл приводится к виду:

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

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. – T. 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. – T. 27. – №. 3. – C. 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