Байесовские методы. Лекция 4.
Байесовский вывод для линейной регрессии.

Целищев М.А.

МГУ им. М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра математической статистики

весна 2021

Список литературы

- J. Kruschke. Doing Bayesian Data Analysis, Second Edition. A Tutorial with R, JAGS, and Stan. Academic Press, 2014.
- R. McElreath. Statistical Rethinking. A Bayesian Course with Examples in R and Stan. Chapman and Hall CRC, 2015.
- O. Martin. Bayesian Analysis with Python. Introduction to Statistical Modeling and Probabilistic Programming using PyMC3 and ArviZ, Second Edition. Packt, 2018.
- K. P. Murphy. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
- A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, D. Rubin. Bayesian Data Analysis, Third Edition. CRC Press, 2013.
- S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng. Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, 2011.

Задача регресии

По вектору признаков $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ хотят предсказать значение целевой переменной $t \in \mathbb{R}$.

Пример: предсказание изменения цены на нефть за день (в процентах). Признаками могут быть изменение цены нефти или других товаров/ценных бумаг за прошлые периоды, текущие процентные ставки и пр.

В наличии исследователя имеется тренировочная выборка

$$D_{\mathrm{tr}} = \{(t_i, \mathbf{x}_i)\}_{i=1}^n, \quad t_i \in \mathbb{R},\ \mathbf{x}_i \in \mathbb{R}^d,$$

то есть конечный набор пар (целевая переменная, вектор признаков). В примере с ценой нефти это могут быть уже известные к текущему моменту наблюдения.

Требуется подобрать алгоритм, который по новому вектору признаков будет делать прогноз на целевую переменную.

Линейная регрессия

Обозначим $\mathbf{t} = (t_1, \ldots, t_n)^T \in \mathbb{R}^n$ — вектор целевых переменных, $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n \times d}$ — матрица признаков (строка $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ которой соответствует целевой переменной t_i).

Тогда с точки зрения линейной модели,

$$t_i \approx w_0 + \sum_{j=1}^{d} w_j x_{ij}, \quad i = 1 \ldots n,$$

где $\mathbf{w} = (w_0, w_1, \ldots, w_d)$ — вектор параметров (весов), который нужно подобрать по тренировочной выборке.

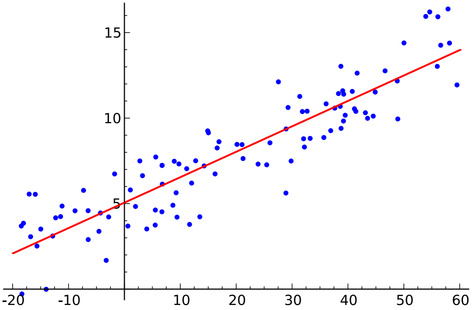
Здесь важно, что целевая переменная *линейно* зависит от *вектора параметров* $\mathbf{w}$, поэтому такая модель и называется *линейной*.

Понятно, что поскольку обычно $n \gg d$, то надеяться на точное равенство не приходится.

Линейная регрессия

Иными словами, хотят провести гиперплоскость в пространстве $\mathbb{R}^{d+1}$, которая бы «наилучшим образом» описывала зависимость между целевой переменной и признаками.

На графике ниже представлен случай $d = 1$.



Линейная регрессия

Для удобства можно считать, что первый признак есть всегда константа 1, то есть $x_{i1} = 1$, $i = 1 \dots n$. В таком случае не нужен w_0 и постановка задачи запишется в виде

$$\mathbf{t} \approx \mathbf{X}\mathbf{w}, \quad \mathbf{t} \in \mathbb{R}^n, \ \mathbf{X} \in \mathbb{R}^{n \times d},$$

и обычно ищут такой вектор весов $\mathbf{w}^* \in \mathbb{R}^d$, который минимизировал бы сумму квадратов ошибок:

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \left(t_i - \sum_{j=1}^{d} w_j x_{ij} \right)^2 = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2$$

Такой инженерный подход называют *методом наименьших квадратов* (МНК).

Как найти оптимальные веса $\mathbf{w}^*$?

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 = \arg\min_{\mathbf{w}\in\mathbb{R}^d} Q(\mathbf{w}),$$

где

$$Q(\mathbf{w}) = \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 = (\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) = \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{t} + \mathbf{t}^T\mathbf{t}.$$

Градиентом числовой функции Q векторного аргумента $\mathbf{w}$ называют вектор частных производных: $\nabla Q(\mathbf{w}) = \left(\frac{\partial Q}{\partial w_1}(\mathbf{w}), \ldots, \frac{\partial Q}{\partial w_d}(\mathbf{w})\right)$.

Поскольку $Q(\mathbf{w})$ — выпуклая функция, то точка её минимума будет совпадать с точкой, в которой градиент равен нулю.

$$\frac{\partial Q}{\partial w_j} = \frac{\partial}{\partial w_j}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}\right) - 2\frac{\partial}{\partial w_j}\left(\mathbf{w}^T\mathbf{X}^T\mathbf{t}\right) =$$

$$= \frac{\partial}{\partial w_j}\left(\sum_{k,m=1}^{d} w_k(\mathbf{X}^T\mathbf{X})_{km}w_m\right) - 2\frac{\partial}{\partial w_j}\left(\sum_{k=1}^{d} w_k(\mathbf{X}^T\mathbf{t})_k\right) =$$

$$= 2(\mathbf{X}^T\mathbf{X})_{jj}w_j + 2\left(\sum_{k\neq j}(\mathbf{X}^T\mathbf{X})_{jk}w_k\right) - 2\left(\mathbf{X}^T\mathbf{t}\right)_j.$$

Линейная регрессия

Итак,

$$\frac{\partial Q}{\partial w_j} = 2\left(\sum_{k=1}^{d}(\mathbf{X}^T\mathbf{X})_{jk}w_k\right) - 2\,(\mathbf{X}^T\mathbf{t})_j.$$

Иными словами,

$$\nabla Q(\mathbf{w}) = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{t},$$

и точкой нулевого градиента будет

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

(при условии, что матрица $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{d\times d}$ обратима).

Матрица $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \in \mathbb{R}^{d\times n}$ называется *псевдообратной* к прямоугольной матрице $\mathbf{X} \in \mathbb{R}^{n\times d}$, и является обобщением понятния обратной для квадратной матрицы.

Линейная регрессия, регуляризация

Что делать, если матрица $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{d\times d}$ необратима? Так будет, если у матрицы $\mathbf{X}^T\mathbf{X}$ есть нулевые собственные значения, что свидетельствует о линейной зависимости столбцов матрицы $\mathbf{X}$.

В таком случае ищут решение регуляризованной задачи:

$$\mathbf{w}_\lambda^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \left(\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2 \right) = \arg\min_{\mathbf{w}\in\mathbb{R}^d} Q(\mathbf{w}, \lambda)$$

для некоторого $\lambda > 0$, где $Q(\mathbf{w}, \lambda) = Q(\mathbf{w}) + \lambda\|\mathbf{w}\|^2$.

Иными словами, накладывается штраф на слишком большие по модулю веса. При этом

$$\nabla_{\mathbf{w}} Q(\mathbf{w}, \lambda) = \nabla_{\mathbf{w}} Q(\mathbf{w}) + 2\lambda\mathbf{w} = 2\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)\mathbf{w} - 2\mathbf{X}^T\mathbf{t},$$

где $\mathbf{I} = \operatorname{diag}(1, \ldots, 1)$ — единичная матрица размера $d \times d$.

Итак, решением регуляризованной задачи будет вектор весов

$$\mathbf{w}_\lambda^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{t},$$

Линейная регрессия, вероятностный подход

А как решают эту же задачу в статистике?

Полагают, что все наблюдения t_i над целевой переменной независимы и имеют нормальное распределение $\mathcal{N}(\mathbf{x}_i^T\mathbf{w}, \sigma^2)$ с некоторой (неизвестной) дисперсией σ^2 и вектором параметров $\mathbf{w} \in \mathbb{R}^d$. Иными словами, вероятностная модель будет иметь вид

$$\mathbf{t} \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}),$$

или, что равносильно,

$$\mathbf{t} = \mathbf{X}\mathbf{w} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}),$$

где ε — шум в наблюдениях, подчиняющийся нормальному закону с нулевым средним и дисперсией σ^2.

Как оценивают неизвестные параметры в статистике?

Линейная регрессия, вероятностный подход

Как всегда, методом максимального правдоподобия!

$$\mathbf{t} \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2 \mathbf{I}),$$

Функция правдоподобия для такой модели есть:

$$\mathcal{L}(\mathbf{t}; \mathbf{w}, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left(-\frac{1}{2\sigma^2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w})\right).$$

Тогда $\arg\max\limits_{\mathbf{w},\sigma} \mathcal{L}(\mathbf{t}; \mathbf{w}, \sigma) = \arg\max\limits_{\mathbf{w},\sigma}\ \ln L(\mathbf{t}; \mathbf{w}, \sigma)$ и

$$\max_{\mathbf{w},\sigma}\ \ln \mathcal{L}(\mathbf{t}; \mathbf{w}, \sigma) = \min_{\mathbf{w},\sigma}\left(\frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2\right).$$

Отсюда следует, что максимизаторами правдоподобия будут:

$$\mathbf{w}_{\mathrm{ML}} = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 = \mathbf{w}^*, \quad \sigma^2_{\mathrm{ML}} = \frac{1}{n}\|\mathbf{t} - \mathbf{X}\mathbf{w}_{\mathrm{ML}}\|^2,$$

т. е. задача максимизации правдоподобия для такой модели эквивалентна МНК.
А как в таком подходе прикрутить регуляризацию???

Линейная регрессия, байесовский подход

Очень просто, нужно провести байесовский вывод!

При этом правдоподобие, как и ранее:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma) \equiv \mathcal{L}(\mathbf{t}; \mathbf{w}, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2\right),$$

то есть

$$(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma) \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}).$$

Стандартный вопрос в байесовском выводе: какое брать априорное распределение для $\mathbf{w}$?

$$p(\mathbf{w}) = p(\mathbf{w}|\mathbf{X}, \sigma) = ???$$

Как водится, надо постараться взять такой класс априорных, чтобы соответствующее апостериорное снова оказалось бы в этом классе (напомним, что такие классы называются *сопряжёнными* к заданному правдоподобию).

Линейная регрессия, байесовский подход

Итак, правдоподобие: $\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma \sim \mathcal{N}(\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I})$, или

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma) \;\propto\; \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2\right)$$

Функционально это квадратичная функция по $\mathbf{w}$ «со знаком минус» под экспонентой. Поэтому в качестве априорного распределения следует брать:

$$\mathbf{w} \sim \mathcal{N}\left(\mu_0, \Sigma_0\right), \quad \mu_0 \in \mathbb{R}^d, \;\Sigma_0 \in \mathbb{R}^{d\times d},$$

то есть $p(\mathbf{w}) \;\propto\; \exp\left(-\frac{1}{2}(\mathbf{w} - \mu_0)^T\Sigma_0^{-1}(\mathbf{w} - \mu_0)\right), \quad \mathbf{w} \in \mathbb{R}^d.$

Тогда по теореме Байеса, апостериорное распределение $\mathbf{w}$ будет иметь вид:

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma) = \frac{p(\mathbf{w}|\mathbf{X}, \sigma)\, p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma)}{p(\mathbf{t}|\mathbf{X}, \sigma)} = \frac{p(\mathbf{w})\, p(\mathbf{t}|\mathbf{w}, \mathbf{X}, \sigma)}{p(\mathbf{t}|\mathbf{X}, \sigma)}.$$

Осталось только посчитать, чему это равно...

Линейная регрессия, байесовский подход

$$p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma) \propto \exp\left(-\frac{1}{2}(\mathbf{w}-\mu_0)^T\Sigma_0^{-1}(\mathbf{w}-\mu_0)\right)\cdot\ \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{t}-\mathbf{X}\mathbf{w}\|^2\right)\propto$$

$$\propto \exp\left(-\frac{1}{2}\left[\mathbf{w}^T\Sigma_0^{-1}\mathbf{w}-2\mu_0^T\Sigma_0^{-1}\mathbf{w}+\frac{1}{\sigma^2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}-\frac{2}{\sigma^2}\mathbf{t}^T\mathbf{X}\mathbf{w}\right]\right)\propto$$

$$\propto \exp\left(-\frac{1}{2}\left[\mathbf{w}^T\left(\Sigma_0^{-1}+\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)\mathbf{w}-2\left(\mu_0^T\Sigma_0^{-1}+\frac{1}{\sigma^2}\mathbf{t}^T\mathbf{X}\right)\mathbf{w}\right]\right)$$

Напомним теперь, как выгдядит плотность нормального распределения $\mathbf{z}\sim\mathcal{N}(\mu,\Sigma)$ в $\mathbb{R}^d$:

$$p(\mathbf{z})\ \propto\ \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right)\ \propto\ \exp\left(-\frac{1}{2}\left[\mathbf{z}^T\Sigma^{-1}\mathbf{z}-2\mu^T\Sigma^{-1}\mathbf{z}\right]\right).$$

Линейная регрессия, байесовский подход

Из сказанного выше делаем вывод, что апостериорное распределение $\mathbf{w}$ имеет вид:

$$\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma \sim \mathcal{N}\left(\mu_n, \Sigma_n\right),$$

где

$$\Sigma_n = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}$$

и

$$\mu_n^T \Sigma_n^{-1} = \mu_0^T \Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{t}^T\mathbf{X},$$

то есть

$$\mu_n^T = \left(\mu_0^T \Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{t}^T\mathbf{X}\right) \cdot \Sigma_n,$$

или

$$\mu_n = \Sigma_n \left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t}\right) = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t}\right).$$

Линейная регрессия, байесовский подход

$$(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma) \sim \mathcal{N}(\mu_n, \Sigma_n)$$

Поскольку максимум плотности нормального распределения достигается в его центре, то оценкой максимума плотности апостериорного распределения (MAP-оценкой) будет апостериорное среднее μ_n:

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}\in\mathbb{R}^d} \; p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma) = \mu_n = \left(\Sigma_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\Sigma_0^{-1}\mu_0 + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t}\right).$$

Если бы в качестве априорных параметров μ_0 и Σ_0 взяли бы

$$\mu_0 = \mathbf{0} \in \mathbb{R}^d, \quad \Sigma_0 = \alpha\mathbf{I} \in \mathbb{R}^{d\times d},$$

то получили бы

$$\mathbf{w}_{\text{MAP}} = \left(\frac{1}{\alpha}\mathbf{I} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{t}\right) = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{t},$$

где $\lambda = \frac{\sigma^2}{\alpha}$. То есть метод наименьших квадратов с регуляризацией есть частный случай байесовского вывода (поиск MAP-оценки для линейной нормальной модели со специальными μ_0 и Σ_0).

Линейная регрессия, байесовский подход

Разобранный подход носит название Ridge-регрессии.

А если бы в качестве априорного взяли *распределение Лапласа*, то есть

$$p(\mathbf{w}) \ \propto \ \exp\left(-\alpha \sum\nolimits_{j=1}^{d} |w_j|\right), \quad \alpha > 0,$$

и решали бы задачу поиска MAP-оценки для $\mathbf{w}$, то снова бы получили

$$\mathbf{w}_{\mathrm{MAP}} = \arg\min_{\mathbf{w}} \left(\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 + \lambda \sum\nolimits_{j=1}^{d} |w_j|\right),$$

но уже с L_1-регуляризатором. Это LASSO-регрессия.

Но байесовский вывод лучше ещё и потому, что позволяет найти не только точечную оценку параметров $\mathbf{w}$, но *целое апостериорное распределение* для них!

Это позволяет оценить степень уверенности или неуверенности в конкретных значениях параметров, строить доверительные интервалы для них, проверять гипотезы и пр.

Линейная регрессия, предсказания

Теперь необходимо сделать предсказание целевой переменной $t_{\text{new}} \in \mathbb{R}$ для нового вектора признаков $\mathbf{x}_{\text{new}} \in \mathbb{R}^d$, имея в наличии тренировочную выборку $D_{\text{tr}} = \{\mathbf{t}, \mathbf{X}\}$.

Классическая статистика даёт такое предсказание на целевую переменную:

$$t_{\text{new}} \sim \mathcal{N}(\mathbf{x}_{\text{new}}^T \mathbf{w}_{\text{ML}}, \sigma_{\text{ML}}^2).$$

А байесовская? Как обычно, надо усреднить правдоподобие нового наблюдения по апостериорному распределению параметров:

$$p(t_{\text{new}}|\mathbf{x}_{\text{new}}, D_{\text{tr}}, \sigma) = \int_{\mathbb{R}^d} p(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma)\, p(\mathbf{w}|D_{\text{tr}}, \sigma)\, d\mathbf{w},$$

где $(t_{\text{new}}|\mathbf{x}_{\text{new}}, \mathbf{w}, \sigma) \sim \mathcal{N}(\mathbf{x}_{\text{new}}^T \mathbf{w}, \sigma^2)$ и $(\mathbf{w}|D_{\text{tr}}, \sigma) \sim \mathcal{N}(\mathbf{w}_{\text{MAP}}, \Sigma_n)$.

ДЗ (*): показать, что

$$(t_{\text{new}}|\mathbf{x}_{\text{new}}, D_{\text{tr}}, \sigma) \sim \mathcal{N}\left(\mathbf{x}_{\text{new}}^T \mathbf{w}_{\text{MAP}}, \sigma^2 + \mathbf{x}_{\text{new}}^T \Sigma_n \mathbf{x}_{\text{new}}\right).$$

Априорное распределение и для σ

В байесовском выводе выше считалось, что шум наблюдений σ известен, но по-байесовски следовало бы навесить априорное распределение и на него.

Если взять априорным

$$\sigma^2 \sim \mathrm{InvGamma}(a_0, b_0),$$
$$\mathbf{w}|\sigma^2 \sim \mathcal{N}(\mu_0, \sigma^2\Sigma_0),$$

то апостериорным тоже окажется гамма-нормальное распределение:

$$\sigma^2|\mathbf{t}, \mathbf{X} \quad \sim \mathrm{InvGamma}(a_n, b_n),$$
$$\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2 \sim \mathcal{N}(\mu_n, \sigma^2\Sigma_n),$$

где $a_n = a_0 + n/2, \quad b_n = b_0 + \frac{1}{2}\left(\mu_0^T\Sigma_0^{-1}\mu_0 + \mathbf{t}^T\mathbf{t} - \mu_n^T\Sigma_n^{-1}\mu_n\right),$

$$\mu_n = \Sigma_n\left(\Sigma_0^{-1}\mu_0 + \mathbf{X}^T\mathbf{t}\right), \quad \Sigma_n = \left(\Sigma_0^{-1} + \mathbf{X}^T\mathbf{X}\right)^{-1}.$$