ЛИНЕЙНАЯ РЕГРЕССИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург
09 сентября 2020 г.

Random facts:

- 9 сентября 9 г. (9.9.9) Публий Квинтилий Вар потерял свои легионы в битве при Тевтобургском лесу; чтобы избежать плена, Вар покончил с собой, но вернуть легионы Октавиану так и не смог
- 9 сентября 1543 г. Мария Стюарт была коронована в Стирлингском замке и стала королевой Шотландии; королеве на тот момент было девять месяцев от роду
- 9 сентября 1937 г. в «Правде» была опубликована «Кантата о Сталине» М.И. Инюшкина: «О Сталине мудром, родном и любимом прекрасную песню слагает народ...»
- 9 сентября 1947 г. Грейс Хоппер прикрепила к своей дневниковой записи моль, которая замкнула цепь в компьютере Mark II; по распространённой версии, отсюда и пошло слово bug в компьютерном смысле, но на самом деле слово гораздо старше, и его употреблял ещё Эдисон
- 9 сентября 2001, в 01:46:40 по Гринвичу, часы отсчитали миллиардную секунду эры UNIX, которая началась в полночь 1 января 1970 года

Линейная регрессия

Метод наименьших квадратов

- Линейная регрессия: рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{x} = (1, x_1, \dots, x_p).$$

- Таким образом, по вектору входов $\mathbf{x}^\top = (x_1, \dots, x_p)$ мы будем предсказывать выход y как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

Метод наименьших квадратов

- Как найти оптимальные параметры $\hat{\mathbf{w}}$ по тренировочным данным вида $(\mathbf{x}_i, y_i)_{i=1}^N$?
- Метод наименьших квадратов: будем минимизировать

$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

- Как минимизировать?

Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

  где $\mathbf{X}$ – матрица $N \times p$, продифференцировать по $\mathbf{w}$, получится

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

  если матрица $\mathbf{X}^\top \mathbf{X}$ невырожденная.
- Замечание: $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$ называется *псевдообратной матрицей Мура–Пенроуза* (Moore–Penrose pseudo-inverse) матрицы $\mathbf{X}$; это обобщение понятия обратной матрицы на неквадратные матрицы.
- Много ли нужно точек, чтобы обучить такую модель?

Байесовская регрессия

- Теперь давайте поговорим о линейной регрессии по-байесовски.
- Основное наше предположение – в том, что шум (ошибка в данных) распределён нормально, т.е. переменная t, которую мы наблюдаем, получается как

$$t = y(x, w) + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

  Иными словами,

$$p(t \mid x, w, \sigma^2) = N(t \mid y(x, w), \sigma^2).$$

- Здесь пока y – любая функция.

Байесовская регрессия

- Чтобы не повторять совсем уж то же самое, мы рассмотрим не в точности линейную регрессию, а её естественное обобщение – линейную модель с базисными функциями:

$$y(x, w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x) = w^\top \phi(x)$$

(M параметров, $M-1$ базисная функция, $\phi_0(x) = 1$).

Байесовская регрессия

- Базисные функции ϕ_i – это, например:
  - результат feature extraction;
  - расширение линейной модели на нелинейные зависимости (например, $\phi_j(x) = x^j$);
  - локальные функции, которые существенно не равны нулю только в небольшой области (например, гауссовские базисные функции $\phi_j(x) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
  - ...

Байесовская регрессия

- Рассмотрим набор данных $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ со значениями $\mathbf{t} = \{t_1, \dots, t_N\}$.
- Будем предполагать, что данные взяты независимо по одному и тому же распределению:

$$p(\mathbf{t} \mid \mathbf{X}, w, \sigma^2) = \prod_{n=1}^{N} N\left(t_n \mid w^\top \phi(x_n), \sigma^2\right).$$

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid w, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - w^\top \phi(\mathbf{x}_n)\right)^2.$$

Байесовская регрессия

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid w, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - w^\top \phi(\mathbf{x}_n)\right)^2.$$

- И вот мы получили, что для максимизации правдоподобия по $\mathbf{w}$ нам нужно как раз минимизировать среднеквадратичную ошибку!

$$\nabla_w \ln p(\mathbf{t} \mid w, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(t_n - w^\top \phi(\mathbf{x}_n)\right) \phi(\mathbf{x}_n).$$

Байесовская регрессия

- Решая систему уравнений $\nabla \ln p(\mathbf{t} \mid w, \sigma^2) = 0$, получаем то же самое, что и раньше:

$$w_{ML} = \left(\Phi^\top \Phi\right)^{-1} \Phi^\top \mathbf{t}.$$

- Здесь $\Phi = (\phi_j(\mathbf{x}_i))_{i,j}$.

Байесовская регрессия

- Теперь можно и относительно σ^2 максимизировать правдоподобие; получим

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - w^\top_{ML}\phi(\mathbf{x}_n)\right)^2,$$

  т.е. как раз выборочная дисперсия имеющихся данных вокруг предсказанного значения.

Спасибо!

Спасибо за внимание!