Прикладная статистика в машинном обучении
Семинар 5
Регрессионный анализ

И. К. Козлов
(*Мехмат МГУ*)

2022

Нелинейная регрессия

Нелинейная регрессия

Почему мы аппроксимируем функцию только линейными?

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i$$

Простейшее обобщение — аппроксимировать линейной комбинацией функций

$$y_i = \beta_0 + f_1(x)\beta_1 + \cdots + f_m(x)\beta_m + \varepsilon_i$$

Подгонка полинома.

Мы хотим приблизить точки (x_i, y_i) полиномом степени $m-1$:

$$y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + \cdots + x_i^{m-1}\beta_{m-1}.$$

Какая будет матрица X?

Подгонка полинома

Подгонка полинома.

Мы хотим приблизить точки (x_i, y_i) полиномом степени $m-1$:

$$y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + \dots + x_i^{m-1}\beta_{m-1}.$$

Какая будет матрица X?

A: Матрица Вандермонда

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^{n-1} \\ 1 & x_2 & \dots & x_2^{n-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^{n-1} \end{pmatrix}.$$

PCA

Главные оси

Q: К какому виду можно привести симметричную билинейную форму

$$B(u,v) = B(v,u), \qquad B(u+v,w) = B(u,w) + B(v,w), \qquad B(\lambda v,w) = \lambda B(v,w)$$

в евклидовом пространстве $\mathbb{R}^n$?

Главные оси

- B — симметричная билинейная форма,
- Q — скалярное произведение (=симметричная положительно определённая билинейная форма),

Приведение пары форм к главным осям:

$$B = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \qquad Q = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}$$

Главные оси

Q: Какой известный объект в Статистике — симметричная билинейная форма?

Главные оси

Q: Какой известный объект в Статистике — симметричная билинейная форма?

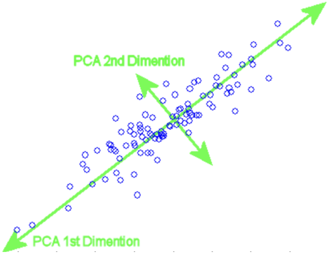
A: Матрица ковариации

$$\mathrm{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

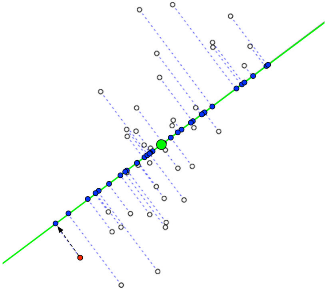
Главные оси

Приводим матрицу ковариации к диагональному виду.

Проецируем данные на оси (главные компоненты).

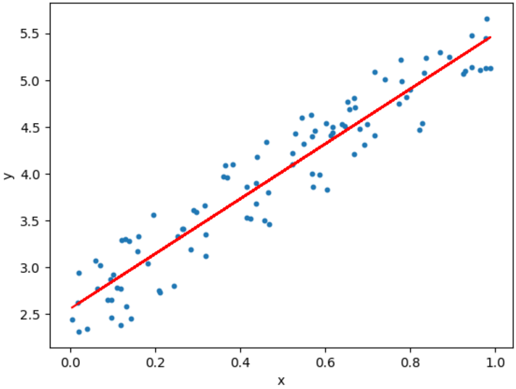


В 2-D проецируем на прямую - 1ую главную компоненту



SVD

Одномерная регрессия — подгонка прямой под облако точек.



Q: Эта прямая — первая главная компонента в PCA или нет?

Дилемма смещения-дисперсия

Разница между одномерной регрессией и PCA:

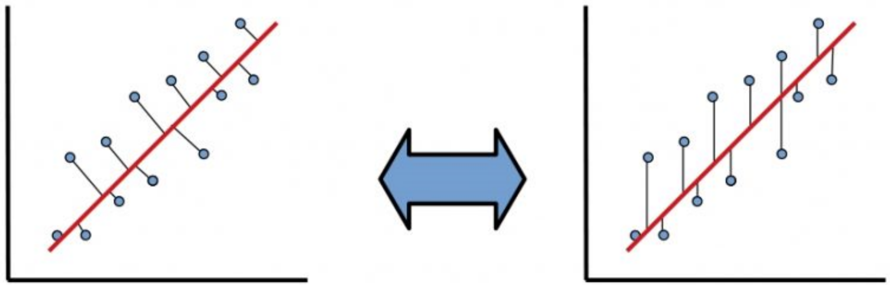


Рис.: PCA vs Linear Regression

Доверительные интервалы

1. Нелинейная регрессия
2. PCA
3. Доверительные интервалы
   - Оценки ошибок
   - Оценки предсказаний

Насколько хороши предсказания регрессия?
Построим доверительные интервалы.

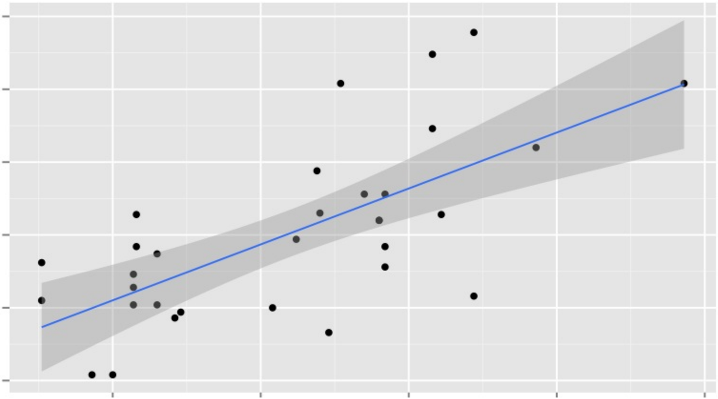


Рис.: Доверительный интервал для регрессии

Гауссовский шум

Гауссовский шум.

Далее будем считать, что шум

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$$

имеет нормальное распределение.

Оценки ошибок

Начнём с оценок для дисперсии ошибок σ^2.

Теорема

Несмещённая оценка для σ^2 — это

$$\hat{\sigma}^2 = \frac{1}{n-m}\,\mathrm{RSS} = \frac{1}{n-m}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Начнём с оценок для дисперсии ошибок σ^2.

Теорема

Несмещённая оценка для σ^2 — это

$$\hat{\sigma}^2 = \frac{1}{n-m}\,\mathrm{RSS} = \frac{1}{n-m}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Более того,

1. RSS независит от оценки $\hat{\beta}$;
2. $\dfrac{1}{\sigma^2}\,\mathrm{RSS} \sim \chi^2_{n-m}$.

Несмещённость

- **Шаг 0**. *Для распределения хи-квадрат* $R_k \sim \chi_k^2$ *матожидание* $\mathbb{E}(R_k) = k$.

  По определению

  $$R_k \sim \chi_k^2 \quad \Leftrightarrow \quad R_k = Z_1^2 + \cdots + Z_k^2, \quad Z_j \sim \mathcal{N}(0,1).$$

  Поэтому

  $$\mathbb{E}(R_k) = \sum_j \mathbb{E}\left(Z_j^2\right) = k\mathbb{V}\left(Z_j\right) = k.$$

Несмещённость

- **Шаг 1**. *Из* $\dfrac{1}{\sigma^2}\,\mathrm{RSS} \sim \chi^2_{n-m}$ *вытекает несмещённость* $\hat{\sigma}^2 = \dfrac{1}{n-m}\,\mathrm{RSS}$.

Несмещённость

- **Шаг 1**. *Из* $\dfrac{1}{\sigma^2}\,\mathrm{RSS} \sim \chi^2_{n-m}$ *вытекает несмещённость* $\hat\sigma^2 = \dfrac{1}{n-m}\,\mathrm{RSS}$.

  Согласно Шагу 0:

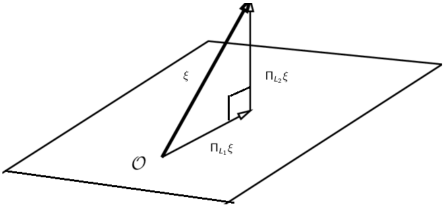
$$\mathbb{E}\left(\frac{1}{\sigma^2}\,\mathrm{RSS}\right) = n - m \quad \Rightarrow \quad \mathbb{E}(\hat\sigma^2) = \sigma^2.$$

Проекция нормального распределения

Вспомним свойство многомерного нормального распределения.

- Пусть случайный вектор $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$.

- L_1 и L_2 — ортогональные подпространства в $\mathbb{R}^n$.



Лемма 1

1. Проекции $\Pi_{L_1}\xi$ и $\Pi_{L_2}\xi$ независимы и нормально распределены.

2. $\dfrac{1}{\sigma^2}\|\Pi_{L_i}\xi\|^2 \sim \chi^2_{\dim L_i}$.

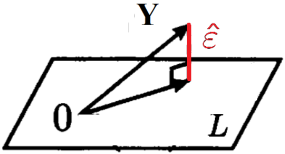
Распределение ошибок

- **Шаг 2.** $\dfrac{1}{\sigma^2}\,\mathrm{RSS} \sim \chi^2_{n-m}$.

Распределение ошибок

- **Шаг 2**. $\frac{1}{\sigma^2}\text{RSS} \sim \chi^2_{n-m}$.

Обозначим $L = L(X)$.



Вектор остатков — это проекция на $L^{\perp}$:

$$\hat{\varepsilon} = \Pi_{L^{\perp}} Y.$$

Распределение ошибок

По условию $Y = X\beta + \varepsilon$. Вектор $X \in L$, поэтому

$$\hat{\varepsilon} = \Pi_{L^\perp} Y = \Pi_{L^\perp} \varepsilon.$$

По Лемме 1

$$\frac{1}{\sigma^2}\,\mathrm{RSS} = \frac{1}{\sigma^2}\left\|\hat{\varepsilon}\right\|^2 \sim \chi^2_{n-m}.$$

Шаг 2 доказан.

Независимость

- **Шаг 3**. RSS *независит от оценки* $\hat{\beta}$.

  Есть два разложения Y:
  $$Y = X\hat{\beta} + \hat{\varepsilon} = X\beta + \varepsilon.$$

  Рассмотрим проекции Y на L и $L^{\perp}$.
  $$\Pi_{L^{\perp}} Y = \hat{\varepsilon} = \Pi_{L^{\perp}}\varepsilon.$$
  $$\Pi_L Y = X\hat{\beta} = X\beta + \Pi_L\varepsilon.$$

  По Лемме 1 они независимы. Получаем, что
  $$\hat{\beta} = \beta + (X^T X)^{-1}\Pi_L\varepsilon.$$

  $\hat{\beta}$ и $\hat{\varepsilon} = \Pi_{L^{\perp}}\varepsilon$ независимы как функции от независимых случайных величин.

Оценки предсказаний

Оценки предсказаний

Пусть мы построили модель регрессии

$$\hat{r}(x) = \hat{\beta}_0 + x\hat{\beta}_1$$

по данным $(x_1, Y_1), \dots (x_n, Y_n)$.

Попробуем предсказать значение Y в новой точке $X = x_*$.

Оценки предсказаний

- Предсказанное значение
$$\hat{Y}_* = \hat{\beta}_0 + x_* \hat{\beta}_1$$
- Истинное значение
$$Y_* = \beta_0 + x_* \beta_1 + \varepsilon.$$

Q: Равны ли дисперсии $\mathbb{V}(\hat{Y}_*)$ и $\mathbb{V}(Y)$?

Оценки предсказаний

- Предсказанное значение

$$\hat{Y}_* = \hat{\beta}_0 + x_* \hat{\beta}_1$$

- Истинное значение

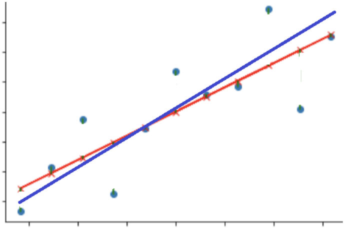
$$Y_* = \beta_0 + x_* \beta_1 + \varepsilon.$$

Q: Равны ли дисперсии $\mathbb{V}(\hat{Y}_*)$ и $\mathbb{V}(Y)$?

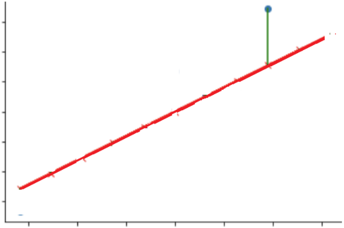
A: Нет, не учитывается шум ε:

$$\mathbb{V}(Y) = \mathbb{V}(\hat{Y}_*) + \sigma^2.$$

При предсказании $\hat{Y}_* = X_* \hat{\beta}$ мы допускаем 2 ошибки:



Параметры $\hat{\beta} \neq \beta$



Есть шум: $Y_* = X_* \beta + \varepsilon$

Оценки предсказаний

Отсюда возникает поправка в доверительном интервале:

13.11 Theorem (Prediction Interval). *Let*

$$\widehat{\xi}_n^2 = \widehat{\sigma}^2 \left(\frac{\sum_{i=1}^n (X_i - X_*)^2}{n \sum_i (X_i - \overline{X})^2} + 1 \right). \tag{13.16}$$

An approximate $1 - \alpha$ *prediction interval for* Y_* *is*

$$\widehat{Y}_* \pm z_{\alpha/2} \widehat{\xi}_n. \tag{13.17}$$