ФКН

Департамент больших данных и информационного поиска

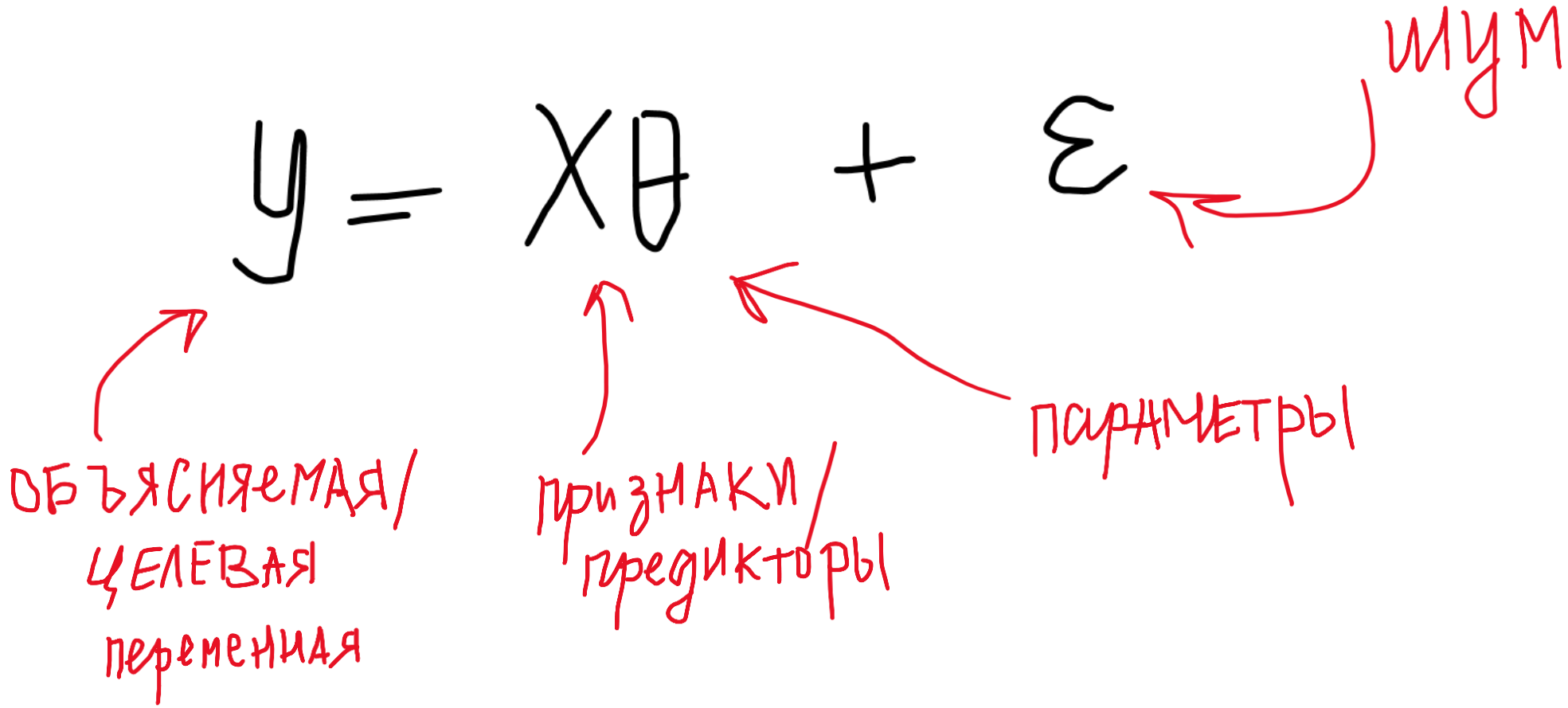
Москва 2025

Лекция 7
Линейная регрессия +

Машинное обучение в цифровом продукте

Полякова И.Ю.

Напоминание



Напоминание

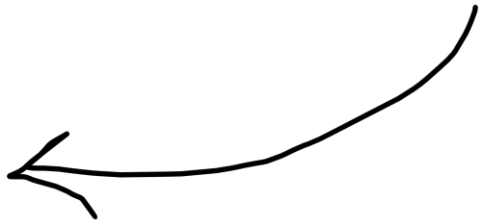
“Линейная регрессия работает всегда, кроме тех случаев, когда она не работает, что происходит почти всегда”

© Комментатор с Youtube

- Выполнены условия Г-М
- Оцениваем МНК
- Получаем **BLUE**

Целевая переменная имеет распределение даже близко, не напоминающее Гауссовое

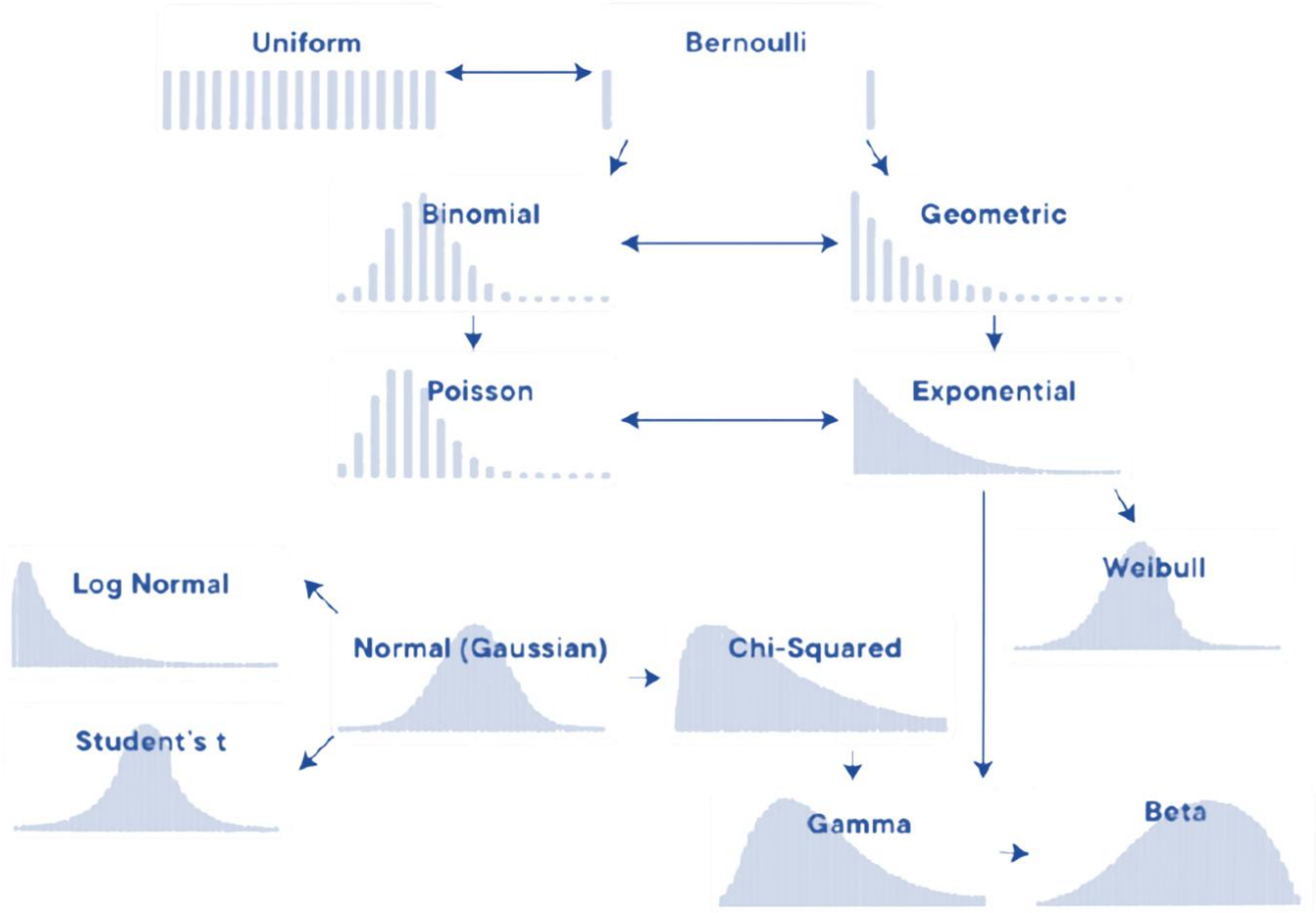
Ошибки автоматически тоже имеют распределение, отличное от нормального



Оценки остаются несмещенными и состоятельными, но перестают быть эффективными

Нельзя быть уверенными в доверительных интервалах и результатах тестов

Целевая переменная имеет распределение даже близко, не напоминающее Гауссовое



Источник: [GLM — Dropbox]

Обобщенные линейные модели

Generalized linear models (GLM)

Идея: введем монотонную и дифференцируемую функцию связи (g)

$$g(y_i) = \theta_0 + \theta_1 x_{i1} + \dots + \theta_k x_{ik}$$

$$\hat{y}_i = g^{-1}(\hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \dots + \hat{\theta}_k x_{ik})$$

Функция связи преобразует распределение целевой переменной так, что :

- Оно принимает значения от $-\infty$ до $+\infty$
- Оно линейно зависит от регрессоров модели

Обобщенные линейные модели

Запишем немного по-другому:

$$y \sim \text{Exponential Family}(\mu, \varphi)$$

здесь вся случайность

$$E(y) = \mu$$

$$g(\mu) = X\theta$$

Или:

$$E(y|X) = \mu = g^{-1}(X\theta)$$

- В модели GLM нет компоненты «случайного шума», так как мы уже зашили его в предположение о распределении переменной;
- В этом смысле постановка 1: ошибки распределены нормально, и постановка 2: целевая переменная распределена нормально, используем функцию связи Identity, - являются эквивалентными

Обобщенные линейные модели

- Оценка GLM позволяет ослабить условия Гаусса-Маркова, а именно отказаться от предпосылки о гомоскедастичности и отсутствии автокорреляции;

- Иначе говоря, ковариационная матрица ошибок может быть любой (естественно, симметричной и положительно определенной, иначе, это не ковариационная матрица)

Обобщенные линейные модели

- Оценка МНК применима, но не гарантирует BLUE;
- GLM зачастую оцениваются с помощью метода максимального правдоподобия;
- Если решение в аналитическом виде для функции правдоподобия недоступно, то используют итеративные методы, в частности IRLS (Iteratively Reweighted Least Squares)

Robust Regression

$$y = X\theta + \varepsilon \qquad \varepsilon \sim Laplace(0, a)$$

Или эквивалентно:

$$y \sim Laplace(\mu, a)$$

$$g = Identity()$$

Функция правдоподобия:

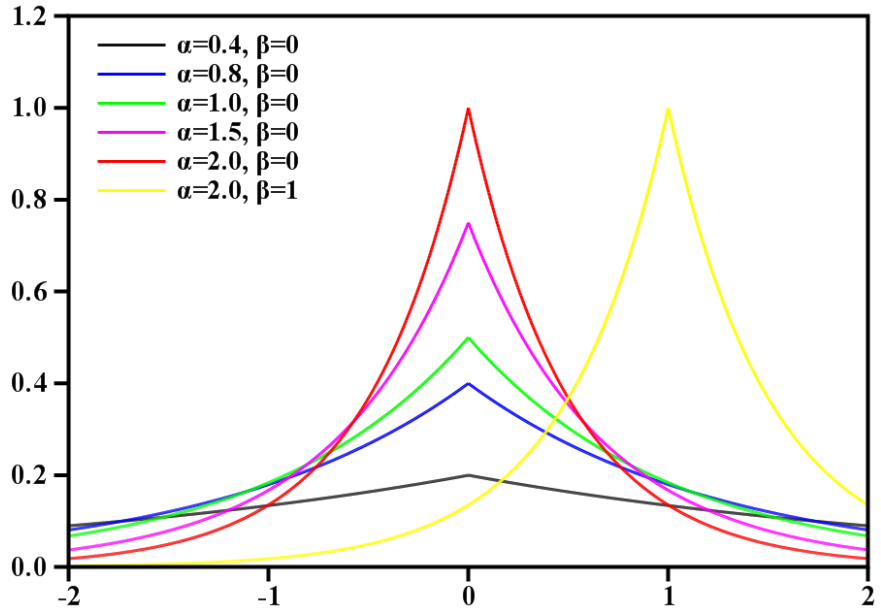
$$L = \left(\frac{a}{2}\right)^n \cdot \exp^{-a \sum |y_i - \theta x_i|}$$

$$\ln L = n \ln \frac{a}{2} - a \sum |y_i - \theta x_i| \to \max_\theta$$

$$\Updownarrow$$

$$\sum |y_i - \theta x_i| \to \min_\theta$$

MAE



$$f(x) = \frac{a}{2} \cdot \exp^{-a|x-\mu|}$$

Robust Regression

$$MAE = \frac{1}{n}\sum |y_i - \hat{y}_i|$$

$$\int |y-\hat{y}| \cdot f(y|x)\,dy \to \min_{\hat{y}}$$

$$\int_{\hat{y}>y} (\hat{y}-y)\cdot f(y|x)\,dy + \int_{\hat{y}<y} (y-\hat{y})\cdot f(y|x)\,dy \to \min_{\hat{y}}$$

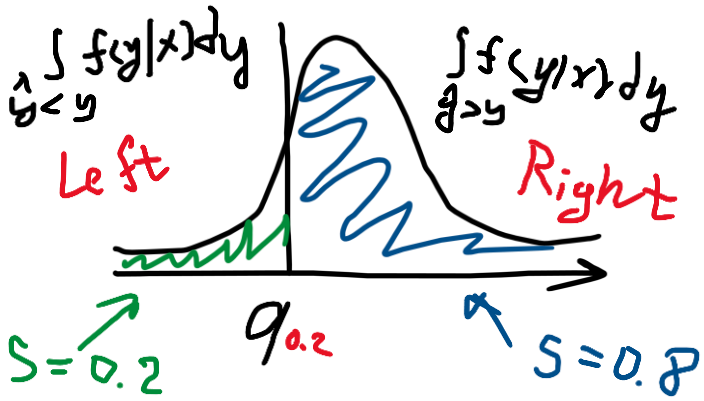
$$\text{FOC}: \frac{d}{d\hat{y}} = \int_{\hat{y}>y} f(y|x)\,dy - \int_{\hat{y}<y} f(y|x)\,dy = 0 \Rightarrow$$

$$\Rightarrow \boxed{\hat{y} = \text{Med}(y|x)}$$

Quantile Regression

Идея: обобщим прошлую концепцию на произвольный квантиль

Пример: хотим оценить квантиль уровня 0.2



Тогда, в финальном результате операций с прошлого слайда, мы бы хотели видеть равенство:

$$\frac{Left}{Right} = \frac{0.2}{0.8} \quad \text{или} \quad Right = 4\,Left$$

В общем виде было бы:

$$\frac{Left}{Right} = \frac{\tau}{1-\tau}, \text{ где } \tau \text{ - уровень квантили}$$

Quantile Regression

Соответствующая такому выводу функция потерь:

$$\sum \rho_\tau (y_i - \hat{y}_i) \to \min_{\hat{y}}$$

Quantile Loss

$$\rho_\tau(u) = \begin{cases} \tau u, & \text{если } u > 0 \\ (\tau - 1)u, & \text{если } u < 0 \end{cases}$$

$$\hat{y} = q_\tau(y|x)$$

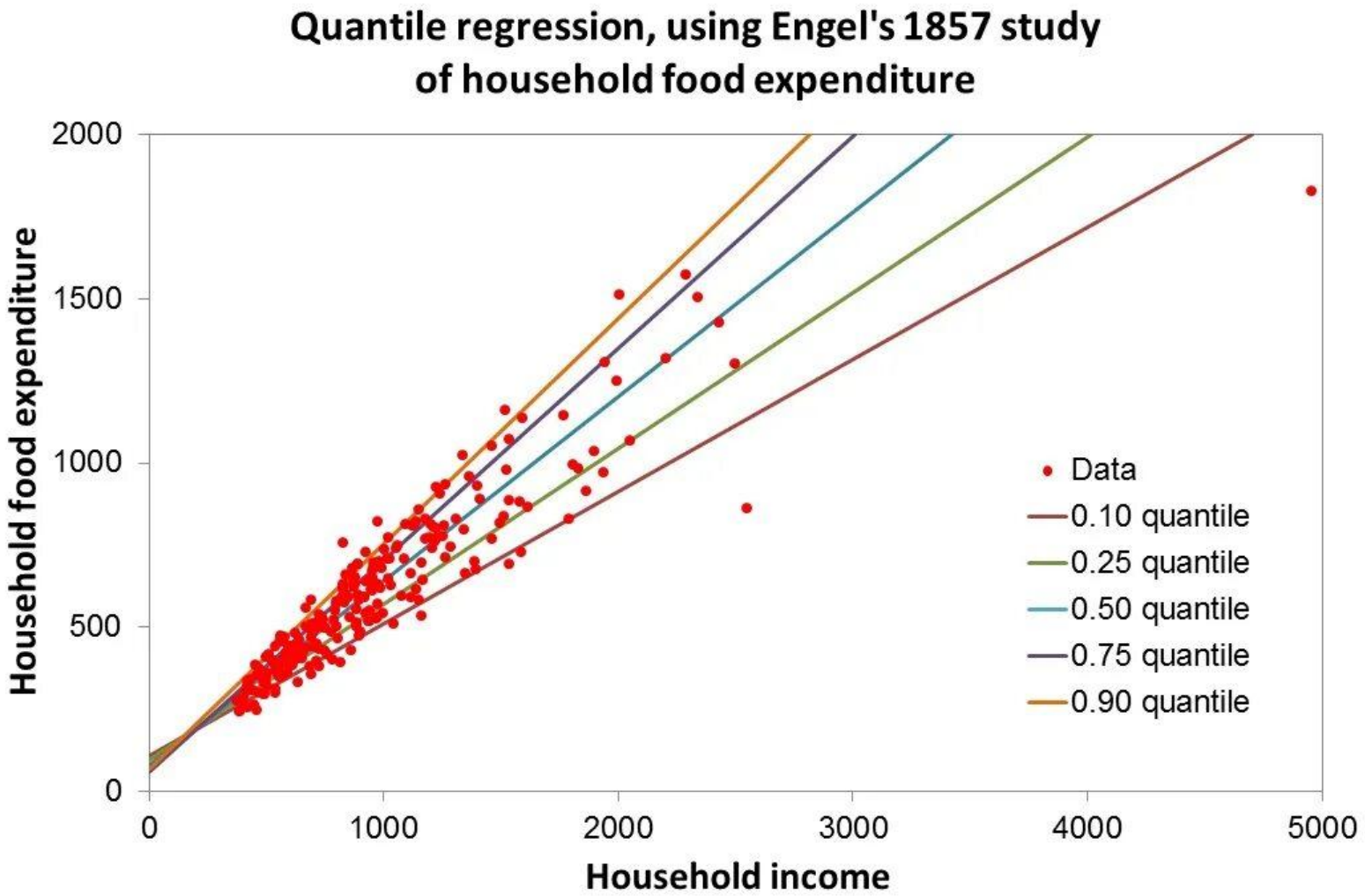
квантиль τ

И исходное распределение ошибок:

$$f(y) = \frac{\tau(1-\tau)}{\sigma} \exp^{-\rho_\tau\left(\frac{y-\mu}{\sigma}\right)}$$

Асимметричная функция Лапласа

Quantile Regression



Logistic Regression

$y_i \sim Bern(p_i)$

$y_i \in \{0, 1\}$ классификация

$$L = p^{\sum y_i} \cdot (1-p)^{n - \sum y_i}$$

$$p_i = P(y_i = 1 \mid X) = g^{-1}(x_i \theta)$$

$p_i(x_i \theta)$

$$L = p_1^{y_1} (1-p_1)^{1-y_1} \cdot \ldots \cdot p_n^{y_n} (1-p_n)^{1-y_n}$$

$$\ln L = \sum_i [y_i \ln p_i + (1-y_i) \ln(1-p_i)] \to \max_\theta$$

$$-\ln L = -\sum_i [y_i \ln p_i + (1-y_i) \ln(1-p_i)] \to \min_\theta$$

LogLoss

Logistic Regression

$$p = \frac{1}{1+\exp^{-x\theta}}$$

sigmoid

$$\ln \frac{p}{1-p} = x\theta$$

При увеличении x на единицу, логарифм отношения шансов увеличивается на θ

$$E[\text{Logloss}\,|x] = \sum_{k \in y} -y \ln p - (1-y)\ln(1-p)\, P(y=k|x) \to \min_p$$

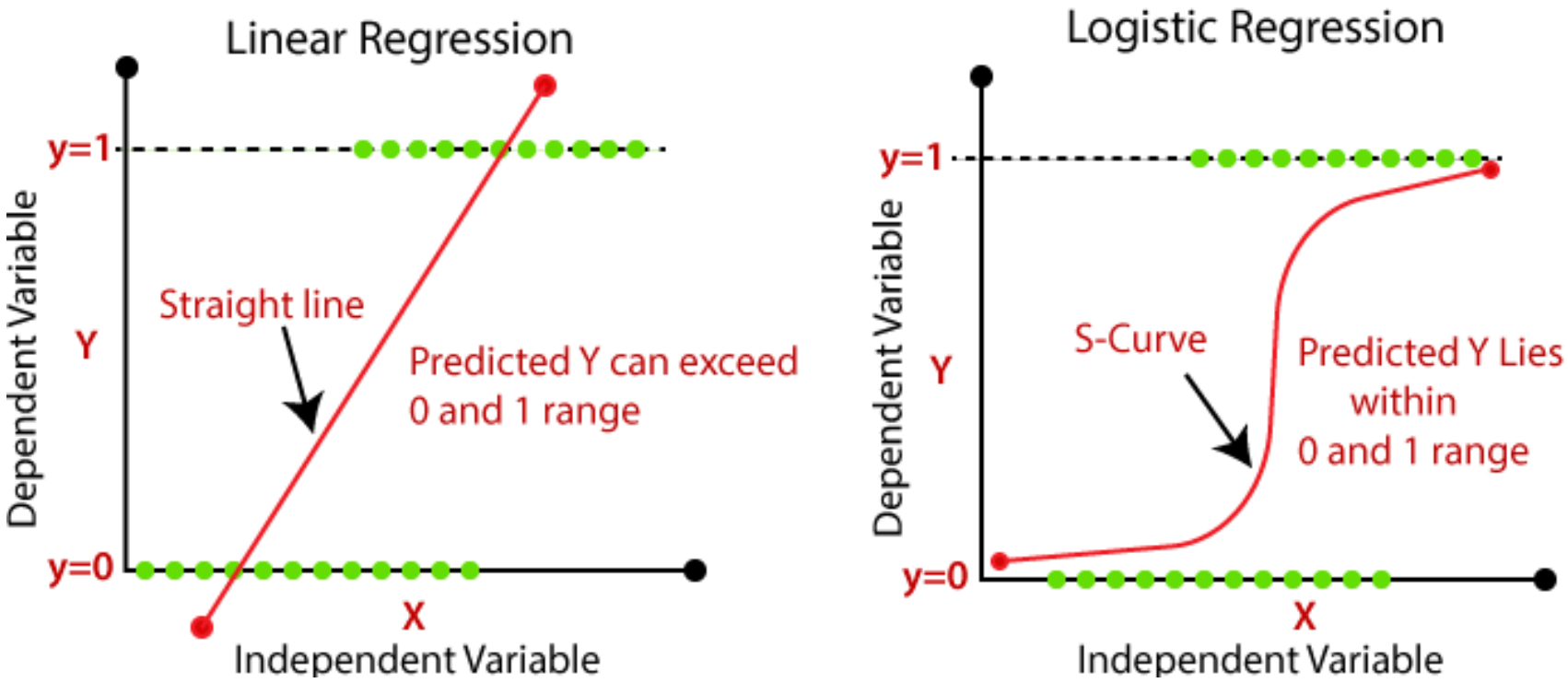
$y = \{0,1\}$ Обозначим: $P(y=1|x) = e$

$$-(1-e)\cdot \ln(1-p) - e \ln p \to \min_p$$

$$\frac{d}{dp} = \frac{1-e}{1-p} - \frac{e}{p} = 0 \Rightarrow (1-e)p = (1-p)e \Rightarrow \frac{p}{1-p} = \frac{e}{1-e} \Rightarrow p = e$$

$p = P(y=1|x)$

Logistic Regression



Источник: https://github.com/Devinterview-io/logistic-regression-interview-questions

Logistic Regression

Как интерпретировать коэффициенты?

- Отношение шансов (Odds Ratio, OR)

При увеличении X на единицу, шансы события увеличиваются в exp^{θ} раз

- Предельный эффект (Marginal Effect, ME)

При увеличении X на единицу, вероятность $P(Y = 1)$ увеличивается на $\theta * p(X) * (1 - p(X))$

$$\frac{dp}{dX_j} = \theta_j \cdot p(X_j) \cdot (1 - p(X_j))$$

То есть, нужно рассчитать вероятность в «базовом» случае, где все предикторы приняли «базовое» значение, а затем подставить в формулу выше

Logistic Regression

Как интерпретировать коэффициенты?

Пример расчета предельных эффектов:

Пусть есть оцененная логит-модель успеха сдачи экзамена по ТВиМС:

$$P(Y=1) = \frac{1}{1 + exp^{-(-9+0.5*\text{часы подготовки})}}$$

$$ME_{\text{Часы подготовки}} = \frac{dP(Y=1)}{d\text{Часы подготовки}} = \frac{exp^{-(-9+0.5*\text{часы подготовки})}}{(1 + exp^{-(-9+0.5*\text{часы подготовки})})^2} * 0.5$$

Мы уже готовились 15 часов, каков будет прирост по вероятности сдачи от дополнительного часа?

$$ME_{\text{Часы подготовки}(15)} = \frac{dP(Y=1)}{d\text{Часы подготовки}(15)} = \frac{exp^{-(-9+0.5*15)}}{(1 + exp^{-(-9+0.5*15)})^2} * 0.5 = 0.07$$

Дополнительный час принесет нам повышение вероятности сдачи на 7%
Если мы посчитаем предельный эффект при условии, что уже готовились 100 часов, он будет близок к 0

Logistic Regression

Как интерпретировать коэффициенты?

Если у нас много предикторов, то придется фиксировать все остальные, кроме того, для которого считаем предельный эффект, на «базовом уровне»

На практике из-за непостоянства предельного эффекта в разных точках принято считать:

- Предельный эффект для среднего по выборке

В нашем примере вычисляем среднее по выборке время подготовки к зачету , а затем считаем предельный эффект для среднего времени

- Средний предельный эффект

В нашем примере вычисляем предельный эффект для каждого студента, затем считаем среднее значение из предельных эффектов

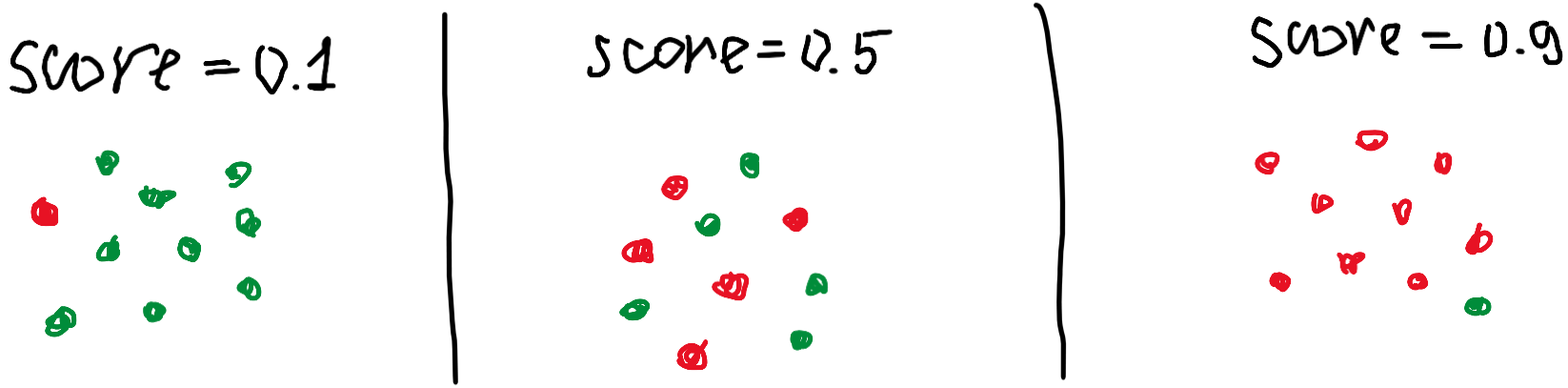
Подробнее: https://books.econ.msu.ru/Introduction-to-Econometrics/chap10/10.2/

Распределения и функции связи

Распределение Y	Функция связи $g(\mu)$	Обратная связь $g^{-1}(\eta)$	Функция потерь в ML	"Классическая" ошибка
Нормальное $N(\mu, \sigma^2)$	μ (identity)	η	**MSE** $\Sigma(y_i - \hat{y}_i)^2$	$\varepsilon \sim N(0, \sigma^2)$
Бернулли $Bern(p)$	**Logit** $\log(p/(1-p))$	**Logistic** $1/(1 + \exp(-\eta))$	**Log Loss** $-\Sigma[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$	Не применимо
Пуассон $Pois(\mu)$	**Log** $\log(\mu)$	**Exp** $\exp(\eta)$	**Poisson Loss** $\Sigma(\mu_i - y_i \log(\mu_i))$	Не применимо
Лапласа $Laplace(\mu, b)$	μ (identity)	η	**MAE** $\Sigma\lvert y_i - \hat{y}_i \rvert$	$\varepsilon \sim Laplace(0, b)$
Асимметричная Лапласа $ALD(\mu, \tau)$	μ (identity)	η	**Quantile Loss** $\Sigma\rho\tau(y_i - \hat{y}_i)$	$\varepsilon \sim AsLaplace(0, \tau)$

Модель предсказывает «вероятность»

Что это значит?



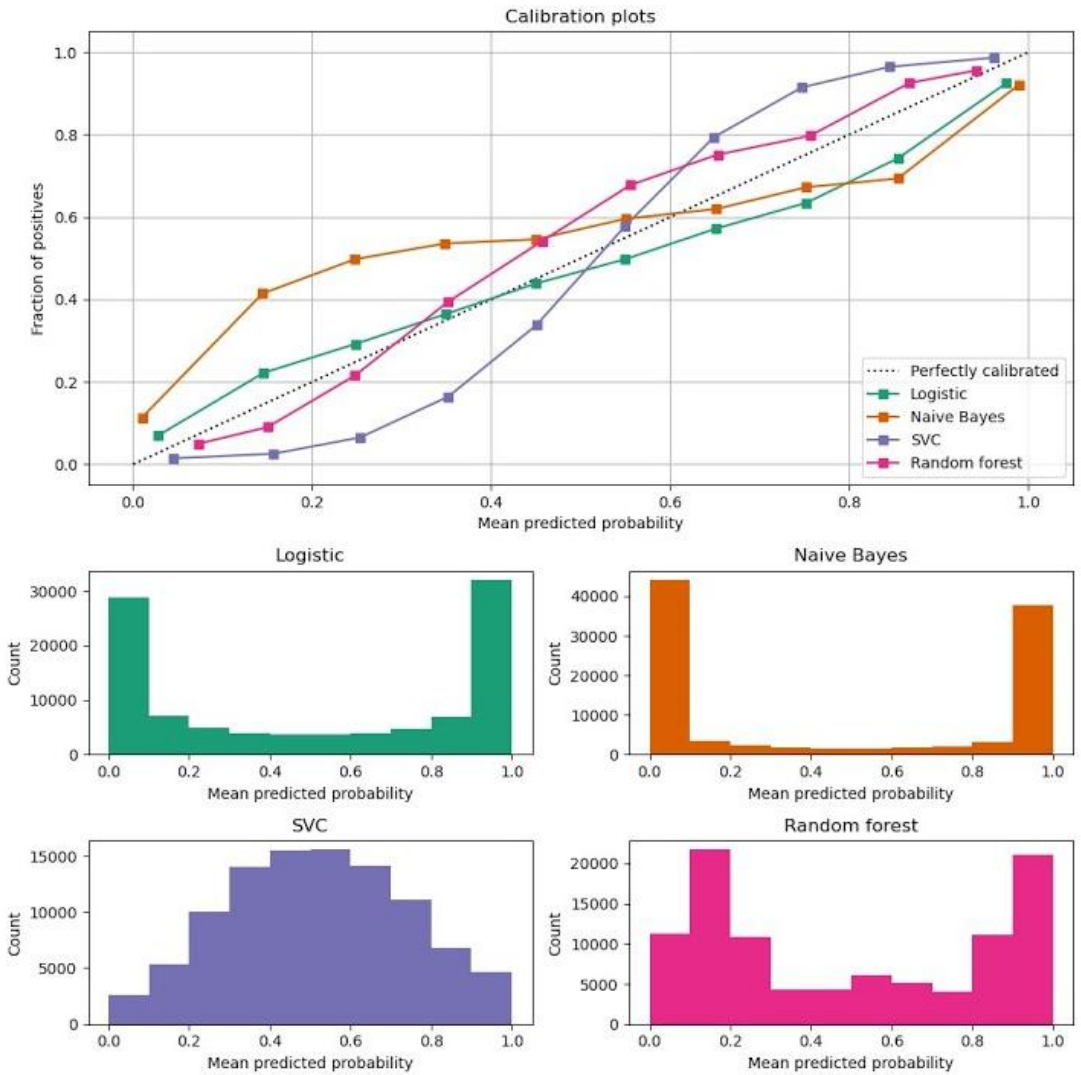
Среди всех объектов, которым модель назначила скор 0.1, доля красных – это 1/10

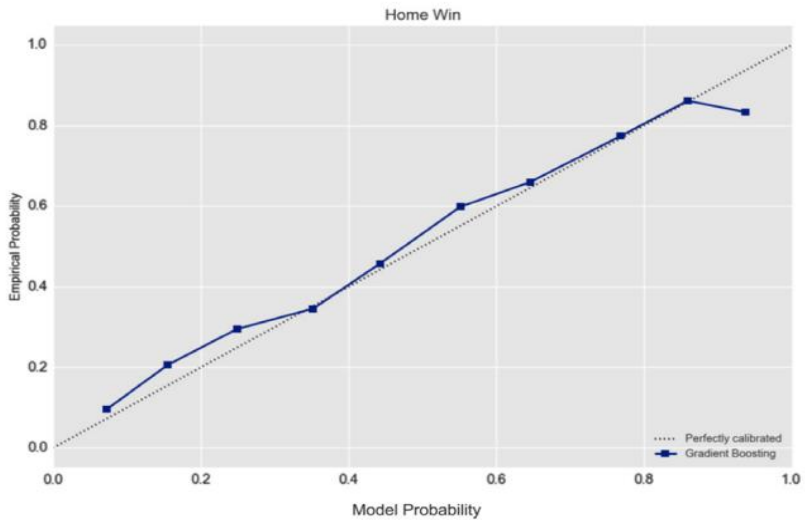
Среди всех объектов, которым модель назначила скор 0.5, доля красных – это 1/2

Среди всех объектов, которым модель назначила скор 0.9, доля красных – это 9/10

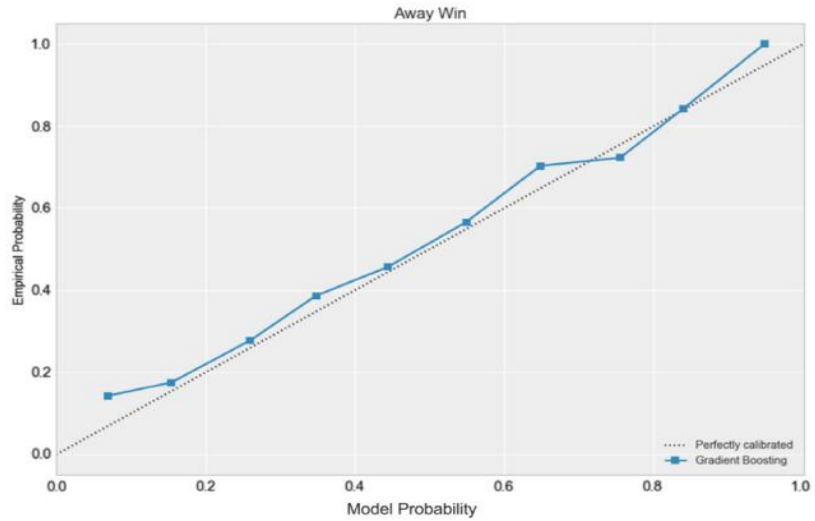
Как доказано на предыдущих слайдах, логит, например, заточен на то, чтобы предсказывать условную вероятность. Однако, меньшинство ML моделей может похвастаться тем же

Калибровочные кривые





(a) Home win.



(b) Away win.

Как калибровать?

Platt Scaling

Идея: построить логит-регрессию на скорах исходной модели (произвольная для бинарной классификации)

$$P_{calib} = \frac{1}{1 + \exp^{-(A \cdot score + B)}}$$

Хорошо работает, если фактическая кривая частот имеет форму, схожую с логистической

Как калибровать?

Isotonic Regression

Идея: “сохраняем порядок” и ищем монотонное преобразование для скоров

Постановка:

Наблюдения: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

$$\begin{cases} \sum_{i=1}^{n} w_i (y_i - f(score_i))^2 \to \min\limits_{f} \\ f(s_1) \le f(s_2) \le \dots \le f(s_n) \end{cases}$$

Хорошо работает со сложными нелинейными искажениями

Подробнее: https://en.wikipedia.org/wiki/Isotonic_regression

Как калибровать?

Isotonic Regression

Алгоритм изотонического восстановления

Идея:

1.Начинаем с тривиального решения: каждая точка — свой сегмент;

2.Объединяем «нарушители»: если два соседних сегмента нарушают монотонность (mean$(A) > mean(B)$), то объединяем их;

3.Повторяем до тех пор, пока весь ряд не станет монотонным.

На выходе имеем кусочно заданную монотонную функцию

Она является наилучшим среднеквадратичным приближением среди всех монотонных функций

Вклад признаков

- Как оценить вклад признаков в линейных моделях достаточно понятно: нужно сравнить коэффициенты в оцененной модели;
- Помним, что признаки нужно привести к одному масштабу, чтобы сравнение было корректным;
- Однако хочется уметь делать что-то подобное для произвольной Ml-модели

Пример

- Пусть есть два друга: Кирилл и Егор, которые хотят немного заработать;
- Они решили петь песни и играть на гитаре, надеясь на хорошие «чаевые»;
- Кирилл играет на гитаре, а Егор поет;
- Известно, что люди неодинаково эмоционально реагируют на музыку без вокала и с ним;
- Если есть только гитара (Кирилл), то он не зарабатывает **ничего**;
- Если есть только голос (Егор), то он зарабатывает **200 рублей**
- Если они **объединяются**: есть и вокал, и аккомпанемент, то они зарабатывают **300 рублей**;
- Как им **справедливо разделить** между друг другом этот выигрыш?

Теория игр: вектор Шепли

Идея: справедливый выигрыш обусловлен тем, какой средний вклад игрок вносит в выигрыш «большой» коалиции, учитывая все возможные варианты ее формирования

Коалиция – объединение k игроков из n, принимающих участие
«Большая» коалиция - коалиция, состоящая из n игроков

Формула компоненты вектора Шепли:

$$\varphi_i(v) = \sum_{i \in k} \frac{(k-1)!\,(n-k)!}{n!} * (v(k) - v(k \backslash \{i\}))$$

Чуть подробнее:
https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D0%BA%D1%82%D0%BE%D1%80_%D0%A8%D0%B5%D0%BF%D0%BB%D0%B8

Теория игр: вектор Шепли

Формула компоненты вектора Шепли:

$$\varphi_i(v) = \sum_{i \in k} \frac{(k-1)!\,(n-k)!}{n!} * (v(k) - v(k \backslash \{i\}))$$

Пояснение:

$\varphi_i(v)$ – выигрыш, который должен достаться i-му игроку

$\sum_{i \in k}$ - сумма по всем коалициям произвольного размера k ($k \leq n$), в которых участвует игрок i

$(k-1)!\,(n-k)!$ – кол-во способов, которыми можно сформировать коалицию размера k, в которой есть игрок i (важно, что игрок i в нее присоединяется последним!)

$n!$ – кол-во способов, которыми можно создать «большую» коалицию (та, в которую включены все игроки)

$(v(k) - v(k \backslash \{i\}))$ - разность выигрыша коалиции размера k с i-м игроком и без него

Теория игр: вектор Шепли

Вернемся к примеру

Все возможные коалиции: {∅}, {Кирилл}, {Егор}, {Кирилл, Егор}

«Большая» коалиция

$$\varphi_K = \frac{(1-1)!(2-1)!}{2!}(0-0) + \frac{(2-1)!(2-2)!}{2!} = 50$$

$$\varphi_E = \frac{(1-1)!(2-1)!}{2!}\cdot(200-0) + \frac{(2-1)!(2-2)!}{2!}\cdot(300-0) = 250$$

Вектор Шепли = (50; 250)

$v\{\varnothing\} = 0$

$v\{\text{Кирилл}\} = 0$

$v\{\text{Егор}\} = 200$

$v\{\text{Кирилл, Егор}\} = 300$

Теория игр: вектор Шепли

Применение в ML

- В качестве игроков – выступают признаки в модели
- В качестве выигрыша – прогноз модели или ее метрика ошибки

- Для формирования коалиций обычно не убирают «лишние» признаки, а заменяют их на случайные значения
- Усреднённый результат модели со случайными значениями признака эквивалентен результату модели, в которой этот признак вообще отсутствует

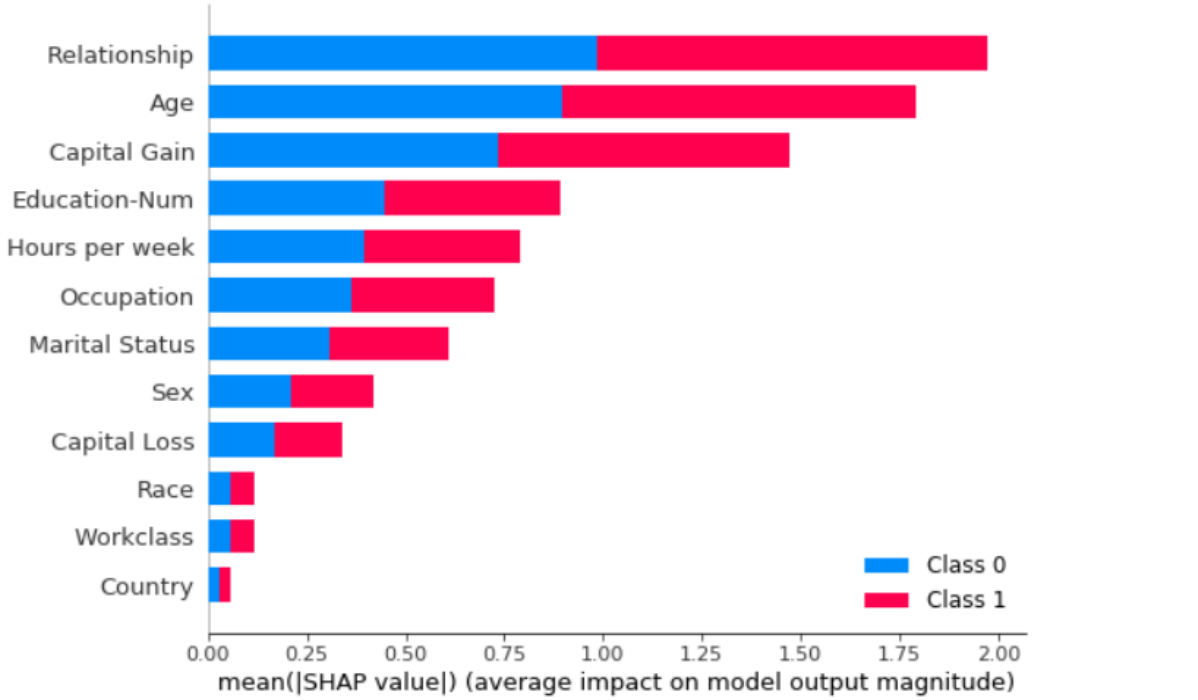
- Конечно, в сложных моделях сложно вычислять вектора Шепли «по-честному» и используют аппроксимацию
- Библиотека SHAP (SHapley Additive exPlanations) поддерживается для моделей типа «ансамбль деревьев» в XGBoost, LightGBM, CatBoost, scikit-learn и pyspark

Библиотека SHAP: https://github.com/shap/shap

SHAP-values

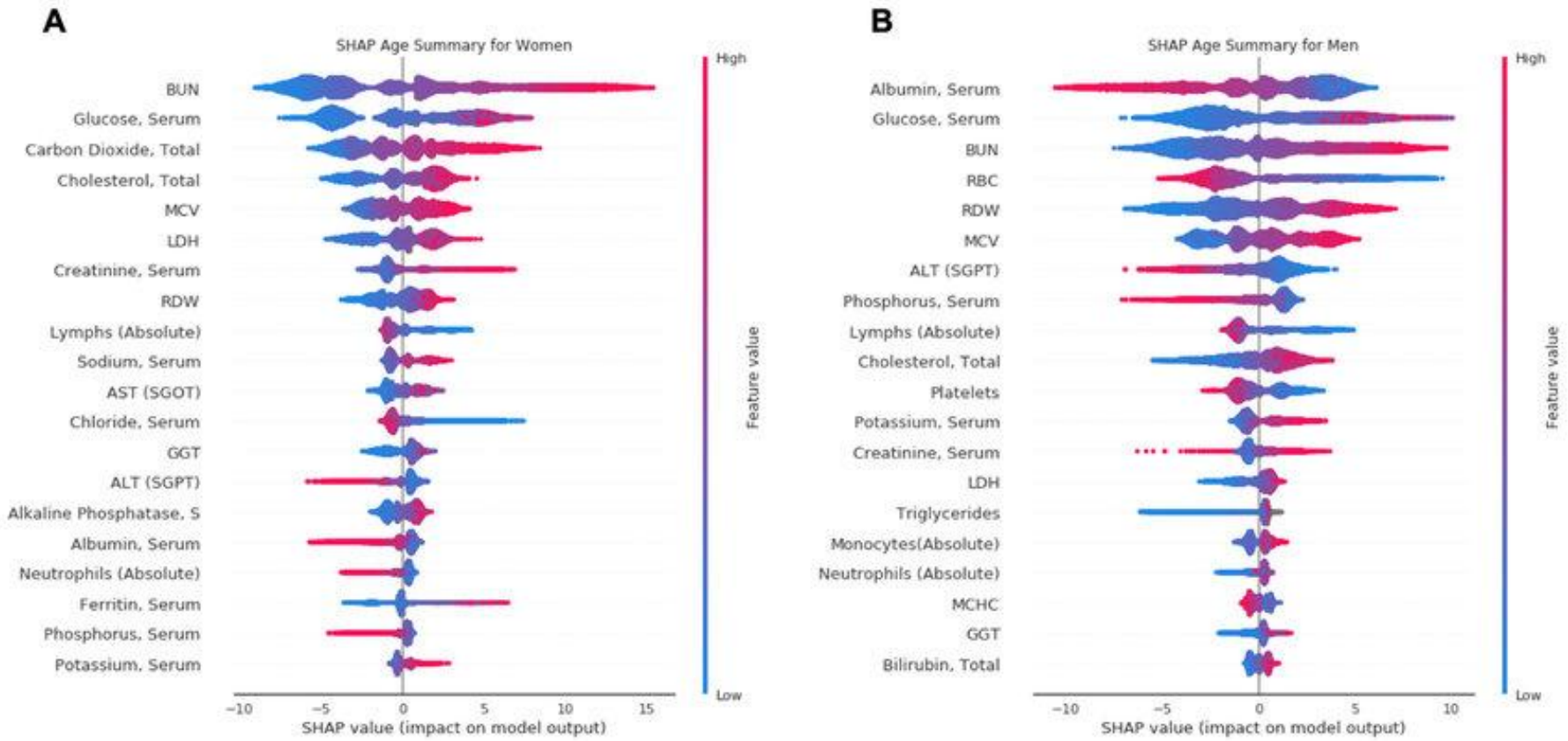
Добавление интерпретируемости в произвольные модели

```
In [7]: shap.summary_plot(shap_values, X)
```



SHAP-values

Добавление интерпретируемости в произвольные модели



Источник: https://www.researchgate.net/figure/SHAP-summary-plots-showing-the-adjustment-to-predicted-age-x-axis-for-each-of-the-top_fig2_330144045

LIME
Local Interpretable Model-agnostic Explanations

Добавление интерпретируемости в произвольные модели

Идея: объясним прогноз сложной модели для интересуемого объекта $\mathbf{x}$ за счёт аппроксимации прогнозов этой модели другой простой и интерпретируемой моделью в окрестности точки $\mathbf{x}$

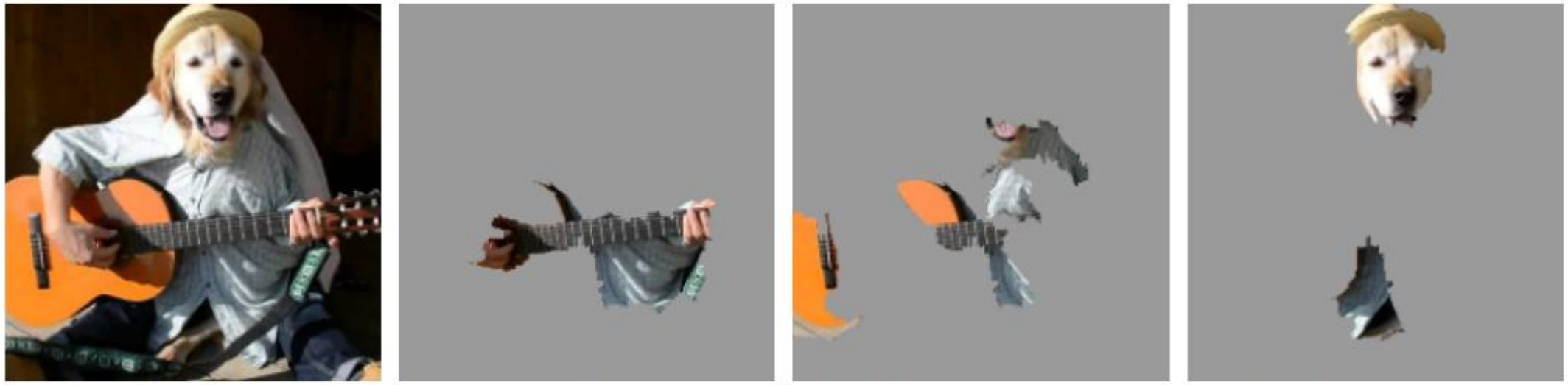
Алгоритм:

- Выбрать объект $\mathbf{x}$, для которого нужно объяснить прогноз сложной модели;
- Сгенерировать выборку, состоящую из локальных вариаций $\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots \tilde{\mathbf{x}}_K$ объекта $\mathbf{x}$;
- Построить для вариаций прогнозы сложной моделью;
- Взвесить объекты по близости к $\mathbf{x}$ (чем вариация ближе, тем её вес больше);
- Обучить простую интерпретируемую модель (чаще всего, линейную регрессию) по взвешенной выборке (чем вес выше, тем объект учитывается сильнее);
- Исследовать веса простой интерпретируемой модели, аппроксимирующей прогноз сложной модели для точки $\mathbf{x}$.

LIME

Local Interpretable Model-agnostic Explanations

Добавление интерпретируемости в произвольные модели



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Источник: https://deepmachinelearning.ru/docs/Machine-learning/Complex-models-interpretation/LIME
Библиотека: https://github.com/marcotcr/lime

Дополнительно

- Про GLM другими словами: https://education.yandex.ru/handbook/ml/article/obobshyonnye-linejnye-modeli
- Хорошее онлайн-пособие по эконометрике, правда про GLM тут не очень подробно: https://books.econ.msu.ru/Introduction-to-Econometrics/chap05/5.5/
- Если интересна именно эконометрика, то очень советую книгу Вербика: https://id.hse.ru/books/1040796649.html (думаю, есть в открытом доступе где-то)
- Про вектор Шепли у Д. Дагаева: Теория Игр. Коалиционные игры 71


ВЫСШАЯ • ШКОЛА • ЭКОНОМИКИ •