CS6462

Probabilistic and Explainable AI

Lesson 12

Bayesian Generalized Linear Models

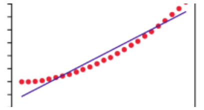
*

Likelihood and Maximum Likelihood Principles

by Emil Vassev

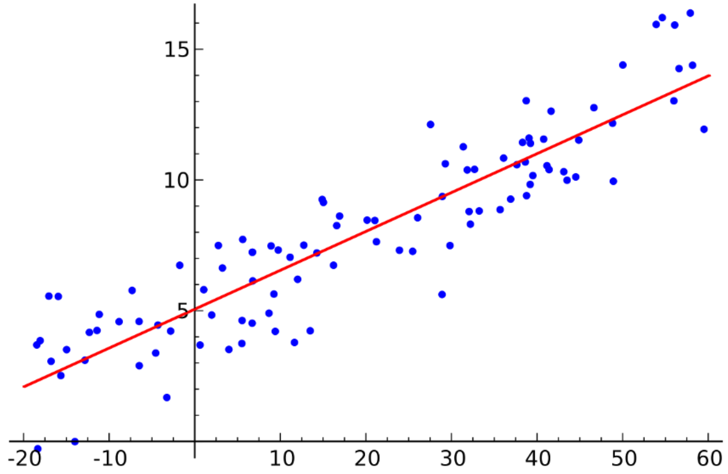
February 24, 2024

Linear Regression



Regression: a statistical model of relationship between a dependent variable and independent variables (one or more) – recall independent variables.

Linear regression: regression with one independent variable and a linear relationship between the independent and dependent variable



$$y = \beta_0 + \ \beta_1 * x$$

- $\boldsymbol{\beta_0}, \boldsymbol{\beta_1}$: scale factor
- **goal**: find the best values for β_0, β_1

$$C(x) = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

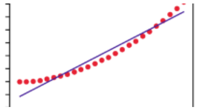
- **Cost function**: difference between the predicted values and real values
- n – number of data points
- Mean Squared Error (MSE)

Predictions for LR: solving the equation for a specific set of inputs (use learning alg.)

Example: predict weight y from height x

β_0 is bias coefficient

- $\boldsymbol{\beta_0}, = 0.1, \boldsymbol{\beta_1} = 0.5$
- $\boldsymbol{x}$ = 182 cm
- **Y = 0.1 + 0.5*182 = 91.1 kg**



Generalized Linear Models

Definition:

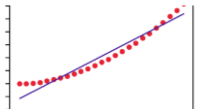
- determines the causal relationship between multiple independent variables and the dependent variable
- the distribution of dependent variables is not only a Normal distribution but can be *Normal, Binomial, Poisson, Categorical, Multinomial, Beta*, etc.

$$y = f(\beta_0 + \beta_1 * x_1 + \beta_2 * x_2 \dots . \beta_n * x_n)$$

$\mathrm{F}(Y) = \mu = g^{-1}(X\beta)$ where $\mathrm{Y} = \{y_1, y_2, \dots, y_k\}, X = \{X_1, X_2, \dots, X_n\}, X_i = \{x_1, x_2, \dots, x_k\}$

- $\mathrm{F}(Y)$ – probability distribution family, e.g., Normal, Binomial, Poisson, Categorical, etc.
- $\eta = X\beta$ – linear predictor, i.e., a linear function of the predictor algorithm (predicted values $pred_i$)
- g – link function:
  - $g(\mu) = \eta$ - maps the expected values μ to the linear predictors $X\beta$
  - $\mu = g^{-1}(\eta)$ - inverse link function (mean function) maps the linear predictors to the mean
  - g and g^{-1} translate η from $(-\infty, +\infty)$ to the proper range for the probability distribution $\mathrm{F}(Y)$ and back again

Maximum Likelihood Estimation



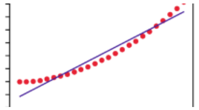
Bayesian GLM: data is fixed, and model parameters are random -> requires prior distribution of parameters

Likelihood theory: part of Bayesian Inference (how data is used by the distribution model)

- Fisher principle: the amount of information that an observable random variable X $(\mathrm{X} \equiv \mathrm{data})$ carries about an unknown parameter $\theta \in \Theta$ $(\Theta \equiv \beta)$ of a distribution that models X
- P(Θ |data) ~ P(data|Θ) * P(Θ) – recall Bayesian Inference for fixed data

  Posterior ~ Likelihood × Prior

Maximum Likelihood Estimator: aims to maximize the probability of every value point of X occurring given a set of probability distribution parameters Θ



Maximum Likelihood Estimation (cont.)

Maximum Likelihood Estimator: aims to maximize the probability of every value point of X occurring given a set of probability distribution parameters Θ

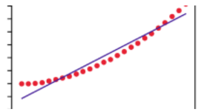
- random sample $\mathrm{X} = \{\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}\}$ with assumed probability distribution $\mathrm{P}(\mathrm{X})$ that depends on some unknown parameter θ
- find linear predictor $\eta(\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n})$ such that $\eta(x_1, x_2, \dots, x_\mathrm{n})$ is an estimate of θ with maximum $\mathrm{P}(\mathrm{X})$ where $(x_1, x_2, \dots, x_\mathrm{n})$ are the observed values of the random sample
- $f(x_i; \theta)$ – the probability density (or mass) function of each X_i
- joint probability (mass) function): likelihood function $\mathbf{L}(\boldsymbol{\theta})$

$$\mathrm{L}(\theta) = \mathrm{P}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = f(x_1; \theta) * f(x_2; \theta) *, \dots, * f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

objective: consider the likelihood function $\mathrm{L}(\theta)$ as a function of θ, and find the value of θ that maximizes it

$$\hat{\theta} = argmax(L(\theta)) \Rightarrow \hat{\theta} = argmax(\ln(L(\theta))) \Rightarrow \ln(\mathrm{L}(\theta)) = \ln\left(\prod_{i=1}^{n} f(x_i; \theta)\right) = \sum_{i=1}^{n} \ln(f(x_i; \theta))$$

- $\hat{\theta}$ - estimate “^”



Maximum Likelihood Estimation (cont.)

Example:

Consider a random sample $\mathrm{X} = \{\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n}\}$ of X employees where

- $\mathrm{X}_i = 0$ if a randomly selected employee who does not own a house
- $\mathrm{X}_i = 1$ if a randomly selected employee who does own a house
- X_i : independent Bernoulli random variables with unknown parameter θ;
- find maximum likelihood estimator of θ, the proportion of employees who own a house

$f(X_i; \theta) = \theta^{X_i}(1-\theta)^{X_i - 1}$, for $X_i = \{0,1\}$

$$\mathrm{L}(\theta) = \mathrm{P}(X_i) = \prod_{i=1}^{n} f(x_i; \theta) = \theta^{X_1}(1-\theta)^{1-X_1} * \theta^{X_2}(1-\theta)^{1-X_2} *, \dots, * \theta^{X_n}(1-\theta)^{1-X_n}$$

Simplifying, by summing up the exponents: $\mathrm{L}(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$

$$\ln(L(\theta)) = \left(\sum x_i\right)\ln(\theta) + \left(n - \sum x_i\right)\ln(1-\theta)$$

find $\hat{\theta} = argmax(\ln(L(\theta)))$

objective: consider the likelihood function L(θ) as a function of θ, and find the value of θ that maximizes it

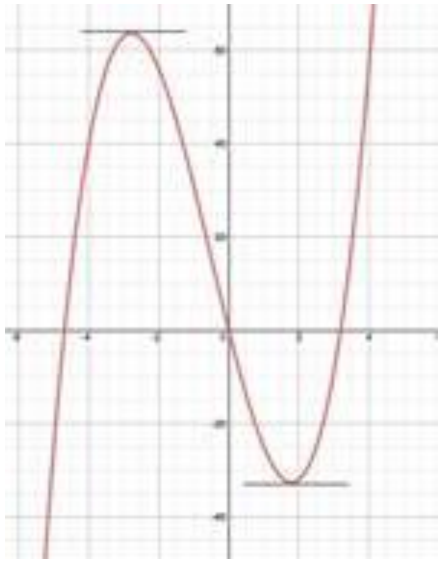
Maximum Likelihood Estimation (cont.)

Example (cont.): objective is to maximize $\mathrm{L}(\theta)$:

find $\hat{\theta} = argmax(\ln(L(\theta)))$

- differentiate the likelihood function with respect to θ
- take the derivative of $\ln(\mathrm{L}(\theta))$ (with respect to θ) rather than taking the derivative of $\mathrm{L}(\theta)$

$$\ln(L(\theta)) = \left(\sum x_i\right)\ln(\theta) + \left(n - \sum x_i\right)\ln(1-\theta)$$



- taking the derivative of the logarithm likelihood, and setting to 0, we get:

$$\frac{\partial \ln(L(\theta))}{\partial \theta} = \frac{\sum x_i}{\theta} - \frac{(n - \sum x_i)}{1-\theta} = 0$$

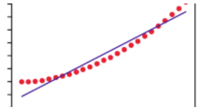
- multiplying through by $\theta(1-\theta)$:

$$\left(\sum x_i\right)(1-\theta) - \left(n - \sum x_i\right)\theta = 0$$

- simplifying: $\boldsymbol{\sum x_i - n * \theta = 0}$

$$\hat{\theta} = \frac{\sum x_i}{n}$$

- $\hat{\theta}$ - estimate “^”
- $\hat{\theta}$ = 1, if all the employees own a house



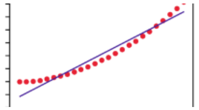
Maximum Likelihood Estimation in ML

Maximum Likelihood Estimation:

- probabilistic technique for solving the problem of density estimation
- involves maximizing a likelihood function in order to find the probability distribution and parameters that best explain the observed data
- provides an approach to predictive modeling in machine learning where finding model parameters can be framed as an optimization problem:

$$\hat{\theta} = argmax(ln(L(\theta;X)))$$

- θ is the parameter space
- X is the observed data (the sample)
- $L(\theta;X)$ is the likelihood of the sample X, which depends on the parameter θ
- argmax function returns the parameter for which the log-likelihood $ln(L(\theta;X))$ attains its maximum value.



Summary

Bayesian Generalized Linear Models - Likelihood and Maximum Likelihood Principles:

- Linear Regression
- Generalized Linear Models
- Bayesian GLM: data is fixed, and model parameters are random -> requires prior distribution of parameters
    - Likelihood Theory: part of Bayesian Inference
    - Maximum Likelihood Estimation: probabilistic technique for solving the problem of density estimation
- Maximum Likelihood Estimation in ML: provides an approach to predictive modeling where finding model parameters can be framed as an optimization problem

Next Lesson:

- Bayesian Generalized Linear Models - Exponential Family Form

Thank You!

Questions?