Lecture 7 - linear & logistic regression

September 24, 2021 3:01 PM

today : linear regression
logistic "

linear regression : derive/motivate with **conditional** approach to regression $(Y \in \mathbb{R})$

$$p(y|x;w) = N(y|\langle w,x\rangle, \sigma^2) \qquad N(\mu,\sigma^2) \quad N(y|\mu,\sigma^2)$$

parameter; $\langle w,x\rangle = w^T x$

$x \in \mathbb{R}^d$
$w \in \mathbb{R}^d$

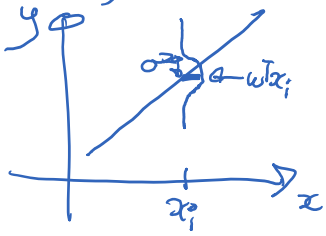
equivalently : $Y_i = w^T X_i + \varepsilon_i$ where $\varepsilon_i | X_i \overset{iid}{\sim} N(0,\sigma^2)$

[aside : we use "offset" notation for x

ie. $x = \begin{pmatrix} \tilde{x} \\ 1 \end{pmatrix}$ $\tilde{x} \in \mathbb{R}^{d-1}$ "constant feature"

thus $\langle w,x\rangle = \langle w_{1:d-1}, \tilde{x}\rangle + w_d$]

"bias" / "offset"



* dataset $(x_i, y_i)_{i=1}^n$ $X_i \sim$ whatever (don't care?)

$Y_i | X_i \overset{indep}{\sim} N(w^T X_i, \sigma^2)$

conditional likelihood

$$p(y_{1:n}|x_{1:n}) \overset{indep.}{=} \prod_{i=1}^n p(y_i|x_i)$$

This not a concave fct. of σ^2

$$\log(\;\;) = \sum_{i=1}^n \left[-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right] \qquad \frac{a}{x} + b\log x$$

$$\frac{\partial}{\partial(\sigma^2)}(\;) = 0 \Rightarrow \sum_{i=1}^n \left[-\frac{(y_i - w^T x_i)^2}{2} \left(\frac{-1}{(\sigma^2)^2}\right) - \frac{1}{2}\frac{1}{\sigma^2} \right] = 0$$

$$\Rightarrow \boxed{\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^n (y_i - w^T x_i)^2}$$

$obj \to -\infty$ as $\sigma \to 0$ or $\sigma \to +\infty$

so conclude that this is correct global max in σ^2 for w fixed

matrix $(\forall i)$

"

(see also

"for w fixed"

(see also note below)

"design matrix" X ($n \times d$ matrix) matrix (▽▽) $\triangleq \begin{pmatrix} - x_1^T - \\ - x_2^T - \\ \vdots \\ - x_n^T - \end{pmatrix}$

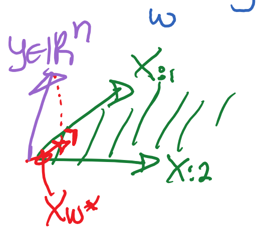
vector $y \triangleq \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

$$Xw = \begin{pmatrix} x_1^T w \\ x_2^T w \\ \vdots \\ x_n^T w \end{pmatrix} \in \mathbb{R}^{n \times 1}$$

$$\sum_{i=1}^n (y_i - w^T x_i)^2 = \|y - Xw\|_2^2$$

can rewrite $-\log p(y_{1:n} \mid X) = \frac{\|y - Xw\|_2^2}{2\sigma^2} + \text{fct}(\sigma^2)$ (design matrix)

MCL $\rightarrow \min_w \|y - Xw\|_2^2 \Longleftrightarrow$ projecting y on the column space of design matrix X



$Xw = \sum_{j=1}^d X_{:j} w_j$ ($X_{:j}$: j^{th} column of X)

$\hat{w}_{MLE} = \arg\min_{w \in \mathbb{R}^d} \|y - Xw\|_2^2$ "least square"

algebra: want $\nabla_w \overset{\text{set to}}{\rightarrow} 0$

$$\frac{\partial}{\partial w}\left[(y - Xw)^T (y - Xw)\right] \overset{\text{want}}{=} 0$$

$\nabla_w (w^T A w)$ (vector) $= (A + A^T) w$

$$\frac{\partial}{\partial w}\left[\|y\|^2 - 2y^T X w + w^T X^T X w\right] = 0$$

(convex fct. of w $\Rightarrow$ stat. pt. is global min)

$$0 - 2X^T y + 2X^T X w = 0$$

$$\Rightarrow \boxed{(X^T X) w^* = X^T y}$$ "normal equations"

a) if $X^T X$ ($d \times d$) is invertible, then have unique sol'n

$$\boxed{\hat{w}_{MLE} = (X^T X)^{-1} X^T y}$$

X is $n \times d$ $\Rightarrow \text{rank}(X) \leq \min\{n, d\}$

$\text{rank}(X^T X) = \text{rank}(X) \leq \min\{n, d\}$

$X^T X$ is invertible $\Rightarrow \boxed{\text{if } n \geq d}$

prediction on training set

$$\hat{y} = X\hat{w} = X\overbrace{(X^T X)^{-1} X^T y}^{\hat{w}_{MLE}}$$

projection matrix on column space of X (recall geometric perspective)

a) if $n < d$ (i.e. high dimension or low data regime) then X^TX is not invertible

b) if X^TX is not invertible → there is no unique soln

any $\hat{w}$ s.t. $(X^TX)\hat{w} = X^Ty$ is a MLE estimate

could choose $\hat{w} = \underset{w:\,(X^TX)w = X^Ty}{\operatorname{argmin}} \|w\|_2 = X^+y$ ← Moore-Penrose pseudo-inverse (see Wikipedia)

$X = \underset{n\times d}{U}\,\underset{n\times d}{\Sigma}\,\underset{d\times d}{V^T}$

$X^+ = V\Sigma^+U^T$

$X^+ = (X^TX)^{-1}X^T$ when X is full rank

$\Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_d \\ & 0 & \end{pmatrix}$ (n rows, d columns) $\quad \Sigma^+ = \begin{pmatrix} \sigma_1^+ & & 0 \\ & \ddots & \\ 0 & & \sigma_d^+ \end{pmatrix}$ (d rows, n columns)

$\sigma_i^+ = \begin{cases} \frac{1}{\sigma_i} & \text{if } \sigma_i > 0 \\ 0 & \text{o.w.} \end{cases}$

problem: pseudo-inverse is not numerically stable

instead it is better to regularize to get similar effect → $\hat{w}_{MAP}(\lambda) \xrightarrow{\lambda \to 0} \hat{w}_{\text{pseudo-inverse}}$

16h00

regularization: (can be motivated from MAP point of view)

Suppose we put a prior $p(w) = \mathcal{N}(w \mid 0, \frac{\sigma^2}{\lambda} I)$ ← I: identity matrix; λ: "precision" parameter

log posterior: $\log p(w \mid \text{data}) = \log p(y_{1:n} \mid X, w) + \log p(w) + \text{cst.}$

$= \frac{1}{2\sigma^2}\|y - Xw\|_2^2 + \text{cst.}(\sigma^2) - \frac{\lambda}{2\sigma^2}\|w\|^2 + \text{cst.}$

MAP here $\hat{w}_{MAP} = \underset{w}{\operatorname{argmin}} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$ ("ridge regression")

same as "regularized" ERM empirical risk min. with squared loss $\ell(y_i, w^Tx_i) = \frac{1}{2}(y_i - w^Tx_i)^2$ → $\frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(y_i, h_w(x_i))}_{\text{squared loss}} + \frac{\lambda}{2n}\|w\|^2$

empirical error; regularization

this is obj. is strongly convex in w ⇒ a unique soln

$f(\cdot)$ is λ-strongly convex

$\Leftrightarrow$ $f(\cdot) - \frac{\lambda\|\cdot\|^2}{2}$ is convex in $(\cdot)$

$$\nabla_w = 0 \Rightarrow (X^TX + \lambda I)w = X^Ty$$

always invertible for $\lambda > 0$

$$\hat{w}_{MAP} = (X^TX + \lambda I)^{-1}X^Ty$$

or ridge regression

no problem for $d > n$

good practice: to either standardize features i.e. make each feature zero mean & unit empirical variance

or normalize $\leftarrow$ make x_i unit norm $\|x_i\|_2 = 1$

or scale features to $[0,1]$ or $[-1,1]$

logistic regression

setup: binary classification $Y = \{0,1\}$ $X \in \mathbb{R}^d$

generative model motivation:

suppose only assumption is there exists a pdf (density) in $\mathbb{R}^d$ for each class conditionals

$$p(x|Y=1) \;\&\; p(x|Y=0)$$

$$p(Y=1|X=x) = \frac{p(Y=1, X=x)}{p(Y=1,X=x) + p(Y=0,X=x)} \;\&\; p(X=x)$$

$$= \frac{1}{1 + \frac{p(Y=0,X=x)}{p(Y=1,X=x)}} = \frac{1}{1+\exp(-f(x))}$$

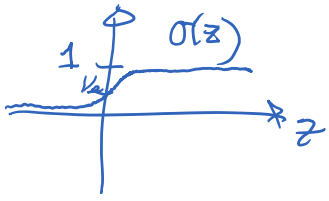
where $f(x) \triangleq \log \frac{p(X=x|Y=1)}{p(X=x|Y=0)} + \log \frac{p(Y=1)}{p(Y=0)}$

"log odds" — class-conditional ratio — prior odd ratio

in general,

$$p(Y=1|X=x) = \sigma(f(x))$$

where $\sigma(z) \triangleq \frac{1}{1+\exp(-z)}$ "sigmoid function"



some properties of $\sigma(z)$:

$$\sigma(-z) = 1 - \sigma(z) \qquad [\sigma(z) + \sigma(-z) = 1]$$

$$\frac{d}{dz}\sigma(z) = \sigma(z)\sigma(-z) = \sigma(z)(1-\sigma(z))$$

(*) to motivate linear logistic regression, consider class conditionals

in the exponential family — scalar fct.: log partition fct. / normalizer

$$p(x|\eta) \triangleq h(x)\exp(\eta^T T(x) - A(\eta))$$

"canonical parameter" — "sufficient statistics" — these specify the "shape" exponential family that we are considering

Gaussian: $\log p(x|\mu,\sigma^2) = -\frac{1}{2}\log 2\pi\sigma^2 - \frac{(x-\mu)^2}{2\sigma^2}$

$\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$

$-\left[\frac{x^2}{2\sigma^2} - x\frac{\mu}{\sigma^2} + \frac{\mu^2}{2\sigma^2}\right]$

let $T(x) = \begin{bmatrix} -x^2/2 \\ x \end{bmatrix}$

$\eta(\mu,\sigma^2) = \begin{bmatrix} 1/\sigma^2 \\ \mu/\sigma^2 \end{bmatrix}$

$A(\eta) = \frac{1}{2}\log|2\pi\sigma^2| + \frac{\mu^2}{2\sigma^2}$

$p(x|Y=1) = p(x|\eta_1)$

$p(x|Y=0) = p(x|\eta_0)$

$$\text{log odds} \quad f(x) = \log \underbrace{\frac{p(x|Y=1)}{p(x|Y=0)}}_{\frac{p(x|\eta_1)}{p(x|\eta_0)}} + \log \underbrace{\frac{p(Y=1)}{p(Y=0)}}_{\frac{\pi}{1-\pi}\triangleq\tau}$$

$$= (\eta_1-\eta_0)^T T(x) + A(\eta_0) - A(\eta_1) + \log\frac{\pi}{1-\pi}$$

$$\triangleq w^T\varphi(x)$$

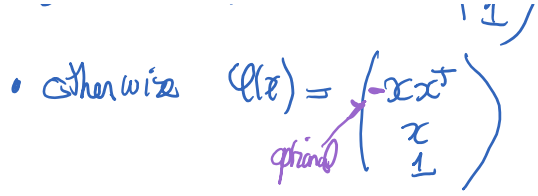
where $w = \begin{pmatrix} \eta_1-\eta_0 \\ A(\eta_0)-A(\eta_1)+\log\frac{\pi}{1-\pi} \end{pmatrix}$ $\quad \varphi(x) = \begin{pmatrix} T(x) \\ 1 \end{pmatrix}$

get logistic regression model $\boxed{p_w(Y=1|X=x) = \sigma(w^T\varphi(x))}$ — "feature map"

decision boundary $\mathbb{1}\{w^T\varphi(x) > 0\}$?

exercise to reader: try argument above with $p(x|y) = N(x|\mu_y, \Sigma_y)$

- if $\Sigma_0 = \Sigma_0$, then $\varphi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix}$
- otherwise $\varphi(x) = \begin{pmatrix} x \\ -xx^T \end{pmatrix}$



- **note about sigma^2 being a global max**

  (**aside**: showing that the sigma^2 above is the **global max** is subtle because the objective is not concave in sigma^2. I give more info here for your curiosity, but it is not required for the assignment.)

  - Formally, to find a global max of a *differentiable objective*, you need to check all **stationary points** (zero gradient points), **as well as the values at the boundary of the domain.**

    Thus here, you would need to show that the objective cannot take higher value anywhere at the boundary of the domain (which is the case here (exercise!), as the objective goes to -infinity at the boundary), so you are done (this is the only possible global optimum -- a maximum here, as it should be, given that there are no other stationary points and all values are lower at the boundary, but one could also explicitly check the Hessian to see that it is strictly negative definite at the stationary point, i.e. it looks like a local maximum).

    Note that we will see later in the class that the Gaussian is in the exponential family, with a log-concave likelihood in the right ("natural") parameterization, and thus using the invariance principle of the MLE, we could also easily deduce the MLE in the "moment" parameterization which is the usual (mu,sigma^2) one, without having to worry about local optima...

  - for a cute counter-example illustrating that a differentiable function could have only one stationary point which is a local min but *not a global min* (and thus why one need to look at the values at the boundary), see:
    - https://en.wikipedia.org/wiki/Maxima_and_minima#Functions_of_more_than_one_variable
    - i.e.

      $$f(x,y) = x^2 + y^2(1-x)^3, \qquad x,y \in \mathbb{R},$$

      shows. Its only critical point is at (0,0), which is a local minimum with $f(0,0) = 0$. However, it cannot be a global one, because $f(2,3) = -5$.

      (see picture of function here)

      (and note that the "Mountain pass theorem" which basically says that if you have a strict local optimum with another point somewhere with the same value, then there must be a saddle point somewhere (a "mountain pass") i.e. another stationary point, **does not hold for this counter-example** as one of the required regularity condition, the "Palais-Smale compactness condition" fails. Here, the saddle point (which should intuitively exist) "happens at infinity", which is why it only has one stationary point despite (0,0) not being a global minimum)
    - the moral of the story: intuitions for multivariate optimization are often misleading! (this counter-example would not work in 1d because of Rolle's theorem)