Lecture 7 - linear & logistic regression

Tuesday, September 26, 2023 2:26 PM

today: linear regression
logistic "

[see lecture last year :]

(asymptotic) properties of MLE:

under suitable regularity conditions on $\mathcal{H}$ & $p(x;\theta)$ $\quad \hat{\theta}_n = \underset{\theta\in\mathcal{H}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log p(x_i;\theta)$

a) $\hat{\theta}_n \xrightarrow{P} \theta$ "consistent"

b) CLT (central limit theorem) $\quad \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d(s)} N(0, I(\theta)^{-1/2})$ $\quad [P_n \sim P_\theta^{\otimes n}]$

information matrix

c) asymptotically optimal (Cramer-Rao lower bound)

ie it has minimal asymptotic scaled variance among all "reasonable" estimators

d) invariance: MLE is preserved under reparameterization

suppose have a bijection $f: \mathcal{H} \to \mathcal{H}'$

then $g(\theta) = f(\hat{\theta})$

example: $(\sigma^2) = (\hat{\sigma})^2$

$\sin \sigma^2 = \sin \hat{\sigma}^2$

* if not a bijection, can generalize the MLE with "profile likelihood"

suppose $g: \mathcal{H} \to \Lambda$

profile likelihood $\triangleq L(\eta) = \max_{\theta:\, \eta = g(\theta)} p(\text{data};\theta)$

define $\hat{\eta}_{MLE} \triangleq \underset{\eta\in\Lambda}{\operatorname{argmax}}\ L(\eta)$

then we have $\boxed{\hat{\eta}_{MLE} = g(\hat{\theta}_{MLE})}$

"plug in estimator"

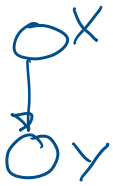
$N(\mu, \sigma^2)$

e.g. $g(\mu) = \mu^2$

Prediction want to learn a prediction fct. $h: \mathcal{X} \to \mathcal{Y}$

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

$\mathcal{Y} = \{0,1\}$ → binary classification
$\mathcal{Y} = \{0, \dots, k-1\}$ → multiclass "
$\mathcal{Y} = \mathbb{R}$ → regression



$$p(x,y) = \underbrace{p(y|x)}_{\text{"prediction model"}} \underbrace{p(x)}_{\text{marginal model over } X}$$

$$= p(x|y)\,p(y)$$

"class conditional" prior over class

"generative perspective" (in context of classifications) → model $p(x)$ as well

vs.

"conditional perspective" → only model $p(y|x)$ (traditionally called "discriminative")

"more discriminative"

generative	conditional	"fully discriminative"
model $p_\theta(x,y)$	model $p_\theta(y\|x)$	model $h_\theta: \mathcal{X} \to \mathcal{Y}$
MLE	max conditional likelihood (MCL)	(not nec. derived from $p(y\|x)$)
		reg. ERM; etc. $\frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, h_\theta(x^{(i)}))$
more assumptions ⇒ less robust predictions		less assumptions more robust

$$\hat{h}(x) \triangleq \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \sum_{y} \hat{p}_\theta(y|x)\, \ell(y, \hat{y})$$

if $\ell(y,\hat{y}) = \mathbb{1}\{y \neq \hat{y}\}$ then $\hat{h}(x) = \operatorname*{argmax}_{\hat{y} \in \mathcal{Y}} p_\theta(\hat{y}|x)$

(0-1 loss)

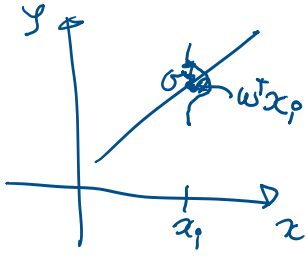
Linear regression: derive/motivate with conditional approach to regression ($Y \in \mathbb{R}$)

$$p(y|x;w) = \mathcal{N}(y \mid \langle w, x \rangle, \sigma^2)$$

w: parameter; $w^T x$

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

equivalently: $Y_i = w^T X_i + \varepsilon_i$



$x \in \mathbb{R}^d$, $w \in \mathbb{R}^d$ 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" (usually denoted as b)

* dataset $(x_i, y_i)_{i=1}^n$ $X_i \sim$ whatever (don't care)
$Y_i | X_i \overset{\text{indep}}{\sim} N(w^T X_i, \sigma^2)$

conditional likelihood:

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

this is not a concave fct. of σ^2

$f(u) = \frac{a}{u} + b \log u$

$$\log(\quad) = \sum_{i=1}^n \left[-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right]$$

want

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

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

obj $\to -\infty$ as $\sigma \to 0$ or $\sigma \to +\infty$ $\Rightarrow$ conclude that this is correct global max in σ^2 for w fixed

(see also notes below)

"design matrix" $X \in \mathbb{R}^{n \times d}$ (matrix ☆☆) $\triangleq \begin{pmatrix} — x_1^T — \\ — x_2^T — \\ \vdots \\ — x_n^T — \end{pmatrix}$ $y \triangleq \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ vector

$$Xw = \begin{pmatrix} x_1^T w \\ \vdots \\ x_n^T w \end{pmatrix} \in \mathbb{R}^{n \times 1} \qquad \sum_{i=1}^n (y_i - w^T x_i)^2 = \| y - Xw \|_2^2$$

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

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

X: design matrix

MCL $\rightarrow \min_w \|y - Xw\|_2^2 \Leftrightarrow$ projection of 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

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

15h32

algebra: want ∇_w set to 0

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

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

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

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

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

$\|\cdot\|^2$ convex fct. of w $\Rightarrow$ stat. pt. is global min

a) if $X^T X$ ($d \times d$) is invertible, then have unique solution

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

prediction on training set

$$\hat{y} = X\hat{w} = \underbrace{X(X^T X)^{-1} X^T}_{\text{projection on column space of } X} y$$

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

$\text{rank}(X^T X) = \text{rank}(X)$

$X^T X$ is invertible $\Rightarrow$ $\boxed{n \geq d}$ $\geq \min\{n, d\}$

(recall geometric perspective)

(*) if $n < d$ (i.e. high dimension or low data regime)

then $X^T X$ is not invertible

[MLE replace with above MCL]

b) if $X^T X$ is not invertible $\rightarrow$ there is no unique sol'n

any $\hat{w}$ s.t. $(X^T X)\hat{w} = X^T y$ is a MCL solution

could choose $\hat{w} = \underset{w}{\text{argmin}} \|w\|_2 = X^\dagger y$ ← Moore-Penrose pseudo-inverse

s.t. $(X^T X) w = X^T y$

lower value $w = \arg\min_w \|w\|_2$ s.t. $(X^TX)w = X^Ty$ — $\hat{w} = X^+ y$ pseudo-inverse

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

SVD

$X = U\Sigma V^T$ $\quad UU^T = I_n \quad VV^T = I_d$

X: $n \times d$, U: $n \times n$, Σ: $n \times d$, V^T: $d \times d$

$\Sigma_{n\times d} = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_d \\ & 0 & \end{pmatrix}$

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

$\Sigma^+ = \begin{pmatrix} \sigma_1^+ & & 0 & \\ & \ddots & & 0 \\ 0 & & \sigma_d^+ & \end{pmatrix}$ $\quad \sigma_i^+ = \begin{cases} 1/\sigma_i & \text{if } \sigma_i \neq 0 \\ 0 & \text{o.w.} \end{cases}$

problem: pseudo-inverse is not numerically stable

instead it is better to regularize to get similar effect.

$\rightsquigarrow \hat{w}_{MAP}(\lambda) \xrightarrow{\lambda \to 0} \hat{w}_{\text{pseudo-inverse}}$

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

suppose we put a prior $p(w) = \mathcal{N}(w \mid 0, \sigma^2 I)$ ($d \times d$ identity matrix)

$\lambda \propto$ "precision parameter"

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

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

MAP here $\hat{w}_{MAP} = \arg\min_{w \in \mathbb{R}^d} \frac{1}{2}\|y - Xw\|^2 + \frac{\lambda}{2}\|w\|_2^2$ "ridge regression"

→ same as "regularized ERM" $\longrightarrow \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(y_i, h_w(x_i))}_{\text{squared loss}} + \underbrace{\lambda}_{\text{regularization}} \|w\|^2$

with squared loss $\ell(y_i, w^Tx_i) = \frac{1}{2}(y_i - w^Tx_i)^2$

→ this obj. is strongly convex in w $\Rightarrow$ a unique solution

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

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

($X^TX + \lambda I$: always invertible $\lambda > 0$)

$\hat{w}_{MAP} = (X^TX + \lambda I)^{-1}X^Ty$ no problem for

$\hat{w}_{MAP} = (X^TX + \lambda I)^{-1}X^Ty$ (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 — make x_i unit norm $\|x_i\|_2 = 1$

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

- **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)