# Lecture 7 - scribbles

Tuesday, September 24, 2019 14:26

today : linear regression
logistic regression

## Prediction

want to learn prediction fct. $h: \mathcal{X} \to \mathcal{Y}$ $\quad \mathcal{Y} = \{0,1\} \to$ binary classification

$x \in \mathbb{R}^d$ $\quad \{0,1,\dots,K\} \to$ multiclass "

$\mathcal{Y} = \mathbb{R} \to$ regression


"prediction model" / model over $\mathcal{X}$

$$p(x,y) = \underbrace{p(y|x)}\ \underbrace{p(x)}$$

$$= \underbrace{p(x|y)}\ \underbrace{p(y)}$$

"class conditional" / prior over classes

"generative perspective" (in context of classification) $\to$ model $p(x)$ as well

"conditional perspective" $\to$ only models $p(y|x)$ ↗ traditionally called "discriminative"

"more discriminative" →

| generative | conditional | "fully discriminative" |
|---|---|---|
| model $p_\theta(x,y)$ | model $p_\theta(y\|x)$ | model $h_\theta : \mathcal{X} \to \mathcal{Y}$ (most rec. derived from $p(y\|x)$) |
| MLE | max. conditional likelihood | |
| more assumptions $\Rightarrow$ less robust for prediction | | less assumptions, more robust |

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

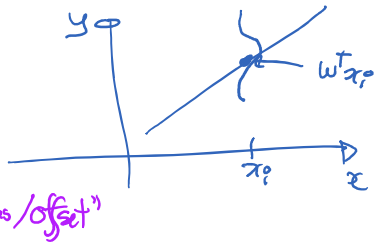
$$p(y|x;w) = N(y|\underbrace{\langle w,x\rangle}_{w^T x}, \sigma^2)$$

$w$: parameter $\quad w \in \mathbb{R}^d$ $\quad x \in \mathbb{R}^d$

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

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

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




[aside: we'll use "offset" notation for x

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

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

- dataset $(x_i, y_i)_{i=1}^n$ $X_i \sim$ whatever

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

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

not a concave fct. of $\sigma^2$ apx

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

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

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

[global min since obj. $\to \infty$ at boundary $\sigma^2 = 0$ or $\sigma^2 \to \infty$]

15h34

$$\left[ -\frac{a}{x} - \log x \right]$$

$$\frac{d}{dx}: \quad \frac{a}{x^2} - \frac{1}{x}$$

$$\frac{d}{dx^2} \quad -\frac{2a}{x^3} + \frac{1}{x^2}$$ ← get sign switch ⇒ not concave

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

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

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

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

($X$: design matrix)

MCL $\rightarrow$ minimizing $\|y - Xw\|^2$ $\Leftrightarrow$ projecting $y$ on the column space of design matrix $X$

(geometric pt. of view)


$$Xw = \sum_{j=1}^{d} X_{:j} w_j$$

$X_{:j}$: $j^{th}$ column of $X$

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

algebra: want $\nabla_w = 0$

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

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

$$\frac{\partial}{\partial w}\left(\|y\|^2 - 2y^T Xw + w^T X^T X w\right) = 0$$

$$\nabla_w \rightarrow \quad 0 - 2X^T y + 2X^T X w = 0$$

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

a) if $X^T X$ is invertible, then have unique solution

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

$X^T X$: $d \times d$ matrix

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

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

prediction on training set $\quad \hat{y} = X\hat{w} = \underbrace{X(X^T X)^{-1} X^T}_{\text{projection operator}} y$

on column space of $X$ (recall geometric

if $n<d$ (ie. high dimensional or low data regime) then $X^TX$ is not invertible

b) what if $X^TX$ is not invertible?

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

Could choose $\hat{w} = \operatorname{argmin}_{w:\, X^TXw = X^Ty} \|w\| = X^+y$ (Moore-Penrose pseudo-inverse)

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

problem: pseudo-inverse is not numerically stable

instead it is better to regularize to get similar effect

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

Suppose we put prior $p(w) = N(w|0, \alpha^2 I)$ ($I$: $d\times d$ identity matrix)

$\lambda$ ← "precision" parameter

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

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

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

same as "regularized" ERM

$\frac{1}{n}\sum_{i=1}^n \ell(y_i, h_w(x_i)) + \frac{\lambda}{2n}\|w\|^2$

squared loss (empirical error) + regularization

is "strongly convex in $w$"

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

$\Rightarrow$ unique solution

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

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

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

no problem for $d > n$

$\hat{w}_{MAP} = (X^TX + \Delta I)^{-1} X^T y$ or ridge regression — no problem for d>n

- **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)