Lecture 7 - scribbles

Tuesday, September 25, 2018 14:31

today: linear regression
logistic regression

Prediction:

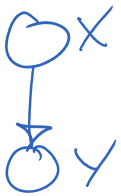
want learn a prediction fct. $h: \mathcal{X} \to \mathcal{Y}$

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

$\mathcal{Y} = \{0,1\}$ → binary classification

$\{0,1,\dots,K\}$ → multiclass "

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



"prediction model" model over X

$$P(X,Y) = \overbrace{P(Y|X)}\,\overbrace{P(X)}$$

prior over classes

$$= \underbrace{P(X|Y)}\,\widetilde{P(Y)}$$

"class conditional"

"generative perspective" → model $p(x)$ as well

"conditional " " → only model $p(y|x)$

↑ traditionally called "discriminative"

more disc. →

gen | conditional | "fully discriminative"

more disc. →

gen	conditional	"fully discriminative"
model $p_\theta(x,y)$	model $p_\theta(y \mid x)$	model $h_\theta : \mathcal{X} \to \mathcal{Y}$ (not nec. $p(y \mid x)$)
MLE	max conditional likelihood	surrogate loss minimization ↳ use $\ell(y, y')$ for estimation
more assumptions ⇒ less robust for prediction		more robust

examples: SVM

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

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

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

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

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

[aside: we'll use "offset" notation for x i.e. $x = \begin{pmatrix} \tilde{x} \\ 1 \end{pmatrix}$ ← "constant feature", $\tilde{x} \in \mathbb{R}^{d-1}$

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

$X_i \sim$ whatever

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

conditional likelihood $p(y_{1:n} \mid x_{1:n}) = \prod_{i=1}^{n} p(y_i \mid x_i)$

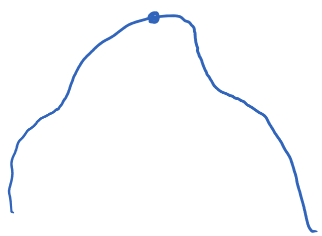
$\log(\quad \text{''} \quad) = \sum_{i=1}^{n} \left[-\frac{(y_i - w^T x_i)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right]$

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

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

$\frac{\partial}{\partial \sigma^2} \rightarrow + \frac{\|y - Xw\|^2}{2v^2} - \frac{1}{2}\frac{1}{v}$

$\frac{\partial^2}{\partial (\sigma^2)^2} \rightarrow - \frac{\|y - Xw\|^2}{v^3} + \frac{1}{2}\frac{1}{v^2}$ & flip sign when v (i.e. σ^2) varies $\Downarrow$ not concave



(see note below about this global max)

"design matrix" $\underset{n \times d \text{ matrix}}{X} \triangleq \begin{pmatrix} — x_1^T — \\ \vdots \\ — x_n^T — \end{pmatrix} \in \mathbb{R}^{n \times d}$ (matrix ($\ddot{w}$))

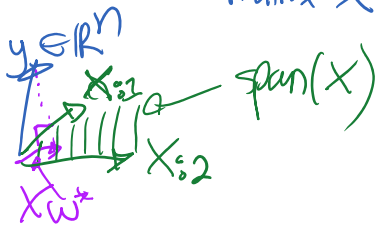
vector: $y \triangleq \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

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

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

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

(geometric view)



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

$X_{:j}$: j^{th} col. of X

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

algebra: want $\nabla_w = 0$

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

$$\frac{\partial}{\partial w}\left(\|y\|^2 - 2y^TXw + w^TX^TXw\right) \overset{\text{want}}{=} 0$$

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

$$= 0 - 2X^Ty + 2X^TXw = 0$$

$$\boxed{(X^TX)w^* = X^Ty}$$ "normal equation"

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

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

X^TX → $d \times d$ matrix
$d \times n \quad n \times d$

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

$X^T X$ → $d \times d$ matrix
$d \times n \quad n \times d$

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

$X^T X$ invertible $\Rightarrow$ $n \geq d$

prediction on training set: $\hat{y} = X\hat{w} = \underbrace{X(X^T X)^{-1} X^T}_{\text{projection operator on column space of } X} y$ (see back geometric viewpoint)

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

15h33

(*) what if $X^T X$ is not invertible?

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

could choose $\hat{w} = \text{argmin}_{w:\, X^T X w = X^T y} \|w\| = X^+ y$ ← Moore-Penrose pseudo-inverse

problem: pseudo-inverse is not numerically stable

instead it is better to regularize to get similar effect

Regularization (motivated from MAP point of view)

suppose we put prior $p(w) = N(w \mid 0, \underbrace{\sigma^2 I}_{\lambda^{-1}})$ ← $d \times d$ identity matrix

λ: "precision" parameter ← multivariate $N(\vec{\mu}, \Sigma_{d \times d})$

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

X, y — w.r.t. to w

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

MAP here

$$\hat{w}_{MAP} = \arg\min_w \frac{1}{2}\|y - Xw\|^2 + \frac{\lambda}{2}\|w\|^2$$

"ridge regression"

same as "regularized" ERM

$$\frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell(y_i, f_w(x_i))}_{\text{squared loss}} + \underbrace{\frac{\lambda}{2n}\|w\|^2}_{\text{regularization}}$$

empirical error

is "strongly convex" in w

$f(\cdot)$ is λ-strongly convex

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

$$\nabla_w = 0 \Rightarrow \underbrace{(X^TX + \lambda I)}_{\text{always invertible } \lambda > 0} w = X^Ty$$

unique solution

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

or ridge regression

no problem for $d > n$

one comment:

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

or

normalize features — make x_i unit norm $\|x_i\|_2 = 1$

or

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

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

logistic regression

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

generative
↳ motivation: suppose only assumption is $\exists$ a pdf (densities) in $\mathbb{R}^d$

$p(x|Y=1)$ & $p(x|Y=0)$ "class conditionals"

$$P(Y=1|X=x) = \frac{P(Y=1, X=x)}{P(Y=1,X=x) + P(Y=0,X=x)} \quad \} P(X=x)$$

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

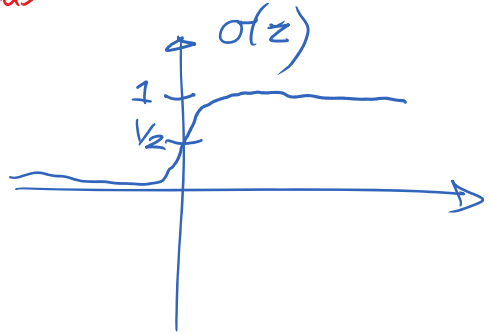
where $f(x) \triangleq \log \underbrace{\frac{P(X=x|Y=1)}{P(X=x|Y=0)}}_{\text{class conditional ratio}} + \log \underbrace{\frac{P(Y=1)}{P(Y=0)}}_{\text{prior odd ratio}}$

"log odds"

and thus in general, $P(Y=1|X=x) = \sigma(f(x))$

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

"sigmoid function"



Some properties of $\sigma(z)$:

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

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

⊛ to motivate linear logistic, consider class conditionals in the exponential family

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

(η: linear in η; "canonical parameter"; $A(\eta)$: log partition fct.; $h(x)$, $T(x)$: scalar functions; these specify the family; $T(x)$: "sufficient statistics")

$$\text{log odds} \quad f(x) = \log \frac{p(x|y=1)}{p(x|y=0)} + \log \frac{P(y=1)}{P(y=0)}$$

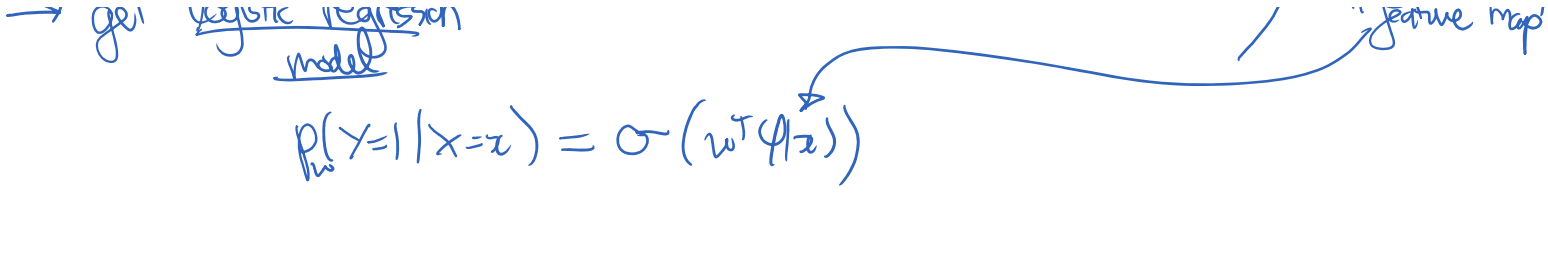
($p(x|y=1) = p(x|\eta_1)$, $p(x|y=0) = p(x|\eta_0)$ parameter for class 0; $P(y=1) = \pi$, $P(y=0) = 1-\pi$)

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

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

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

→ get logistic regression model



- **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)