Lecture 6 - scribbles

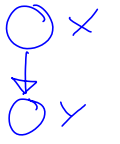
Tuesday, September 20, 2016
14:27

today: gen. vs. discriminative methods
linear & logistic regression

prediction

want to learn a prediction function $f: \mathcal{X} \to \mathcal{Y}$ ($\mathcal{X} \in \mathbb{R}^d$)

$\{0,1\}$ → binary classification
$\{0,1,\dots,K\}$ → multiclass "
$\mathbb{R}$ → regression



$P(X,Y) = P(Y|X)P(X)$ — conditional model on Y; prior over documents
$= P(X|Y)P(Y)$ — class conditional; prior over classes
} generative perspective (also models $p(x)$)

conditional perspective: only models $P(Y|X)$
("traditionally called "discriminative")

gen	conditional	disc. continuum for prediction "fully disc"
model $p(x,y)$	model $p(y|x)$	model $f: \mathcal{X} \to \mathcal{Y}$ (not nec. $p(y|x)$)
more assumptions ⇒ less robust for prediction		more robust

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

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

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

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

equivalently $Y = w^T X + \varepsilon$ where $\varepsilon \overset{\text{indep.}}{\sim} N(0,\sigma^2)$

[aside: we'll use "offset" notation for x $\quad x = \begin{pmatrix}\tilde{x}\\ 1\end{pmatrix}$ $\tilde{x} \in \mathbb{R}^{d-1}$
1 ← constant feature bias/offset]

thus $w^T x = w_{1:(d-1)}^T \tilde{x} + w_d$

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

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

$N(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$

conditional likelihood $p(y_1,\dots,y_n | x_1,\dots,x_n)$
$= \prod_{i=1}^n p(y_i|x_i)$

log-conditional likelihood:

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

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

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

design matrix $X = \begin{pmatrix} \text{---} & x_1^T & \text{---} \\ & \vdots & \\ \text{---} & x_n^T & \text{---} \end{pmatrix} \in \mathbb{R}^{n \times d}$ (column)

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

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

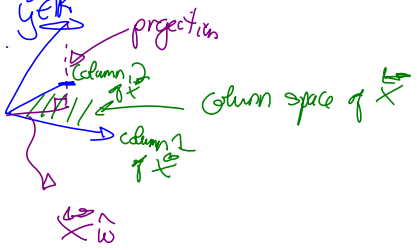
$$\|\vec{y} - Xw\|_2^2 = \sum_{i=1}^{n} (y_i - x_i^T w)^2$$

can rewrite $-\log p(\vec{y} \mid X) = \frac{\|\vec{y} - Xw\|^2}{2\sigma^2} + \text{function}(\sigma^2)$

minimizing $\|\vec{y} - Xw\|^2$ with respect to w → projecting $\vec{y}$ on the column space of X

(geometrical view)

$\vec{y} \in \mathbb{R}^n$



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

j^{th} column of X

algebra:

$$\frac{\partial}{\partial w}(y - Xw)^T(y - Xw)$$

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

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

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

[∇ gradient notation]

constant b: $\nabla_w(w^T A w) = (A + A^T)w$

- if $X^T X$ is invertible, unique solution is

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

maximum conditional likelihood

$X^T X$ (d × n)(n × d) → d × d matrix

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

if $n < d$, then X is not full rank and so $X^T X$ is not invertible

⊗ what if $X^T X$ is not invertible?

(pseudo-inverse would choose the minimum norm $\|w\|$ solution amongst $\arg\min_w \|y - Xw\|$)
↳ $\hat{w} = X^+ y$

problem: pseudo-inverse is not numerically stable

instead better to regularize

MAP estimator viewpoint or the regularized ERM

$\to \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_w(x_i)) + \frac{\lambda \|w\|^2}{2}$ (squared loss; prediction; regularizer; empirical error)

add prior over w: $p(w) = N(w \mid 0, \frac{I}{\lambda})$ (identity in dimension d; λ ? precision)

$N(w \mid \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(w-\mu)^T \Sigma^{-1} (w-\mu)\right)$

$\log p(w \mid y, x) = \log p(y_{1:n} \mid x_{1:n}; w) + \log p(w) + cst.$ (data)

as before; $+ cst. - \frac{\lambda}{2}\|w\|^2$

"ridge regression" [with $\|w\|^2$ penalty]

$\nabla_w = 0 \Rightarrow (X^T X + \lambda I) w = X^T y$

$-\log p(w \mid y; x)$ is strongly convex in w $\Rightarrow$ unique global minimum

$\Rightarrow \hat{w}_{MAP} = (X^T X + \lambda I)^{-1} X^T y$ (or ridge regression; always invertible)

one comment: good practice to standardize features $(x_1, \dots, x_d) \to$ make them empirical zero mean & unit std
or normalize $\to$ put on $[0,1]$ scale or something

Logistic regression

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

no assumption apart that $p(x \mid Y=1)$ & $p(x \mid Y=0)$ are densities

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

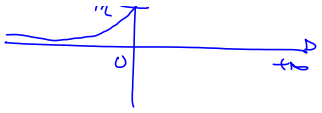
$$= \frac{1}{1 + \frac{p(Y=0, X=x)}{p(Y=1, X=x)}} = \frac{1}{1 + \exp(-f(x))}$$

where $f(x) \triangleq \log \frac{p(X=x \mid Y=1)}{p(X=x \mid Y=0)} + \log \frac{p(Y=1)}{p(Y=0)}$ (class-conditional ratio; prior odd ratio)

in general

$p(Y=1 \mid X=x) = \sigma(f(x))$ where $f(x)$ "log odds"

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



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

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

for example: consider class conditionals in the exponential family

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

η: canonical parameters; $T(x)$: sufficient statistics; $\eta^T T(x)$: linear part in η

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

$p(x|Y=1) = p(x|\eta_1)$, $p(x|Y=0) = p(x|\eta_0)$, $p(Y=1) = \pi$, $p(Y=0) = 1-\pi$

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

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

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

(logistic regression model $p_w(Y=1|X=x) = \sigma(w^T\varphi(x))$)