

### 3 Decision Theory

Friday, October 4, 2019 4:29 PM

<u>Decision Theory</u>		"minimize expected loss"			
<u>Spam</u>	$x, y, \hat{y}$	$\hat{y} = 0$	Spam	$y=0$	$y=1$
Loss fun.		prob	true	Spam	Ham
$L(y, \hat{y})$	Gen. Framework	$\hat{y} = 0$	Spam	0	100
$\epsilon$ :	State $s$ (unknown)	$\hat{y} = 1$	Ham	1	0
	Observation (known)				(loss)
	Action $a$				Reward / Util
	Loss $L(s, a)$				

(ML 3.1) Decision theory (Basic Framework)

"0-1 loss"  $L(y, \hat{y}) = I(y \neq \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{else} \end{cases}$

"Square loss"  $L(y, \hat{y}) = (y - \hat{y})^2$

DT for sup. learning  $(x_1, y_1), \dots, (x_n, y_n) \quad x, y, \hat{y}$

① Given  $x$ , minimize  $L(y, \hat{y})$  ... but don't know  $y$ !

② Choose  $f$  ( $f(x) = \hat{y}$ ) to minimize  $L(y, f(x))$  ...

but don't know  $x$  or  $y$ !

## Decision theory as it applies to Supervised Learning

### 3.2 Minimizing conditional expected loss

Small loss on average  $(x, y) \sim p$

$$\textcircled{1} \quad E(L(Y, \hat{y}) | X=x) = \sum_{y \in Y} L(y, \hat{y}) p(y|x)$$

"0-1 loss"

$$E(L(Y, \hat{y}) | X=x) = \sum_{y \neq \hat{y}} p(y|x) = 1 - p(\hat{y}|x)$$

$$\hat{y} = \arg \min_y E(L(Y, \hat{y}) | X=x) = \boxed{\arg \max_y p(y|x)}$$

### 3.3 Choosing $f$ to minimize expected loss

$$\begin{aligned} \textcircled{2} \quad \hat{Y} &= f(X) \quad EL(Y, \hat{Y}) = EL(Y, f(X)) \\ &= \sum_{x,y} L(y, f(x)) \underbrace{p(x,y)}_{p(y|x)p(x)} = \sum_x \left( \underbrace{\sum_y L(y, f(x)) p(y|x)}_{g(x, f(x))} \right) p(x) \\ &= \sum_x g(x, f(x)) p(x) = E^X g(X, f(X)) \geq E^X g(X, f_o(X)). \end{aligned}$$

Suppose for some  $x', t$ :

$$g(x', f(x)) > g(x', t) \quad f_o(x) = \begin{cases} f(x) & \text{if } x \neq x' \\ t & \text{if } x = x'. \end{cases}$$

$$\forall x \quad g(x, f(x)) \geq g(x, f_o(x))$$

$$\overline{x, y} \rightarrow \overbrace{x, y}^{\text{pair}} \quad x \quad | \quad y \quad \rightarrow \quad \overbrace{x, y}^{\text{pair}}$$

$$\begin{aligned}
 & p(y|x) p(x) \overbrace{g(x, f(x))} \\
 &= \sum_x g(x, f(x)) p(x) = E^X g(X, f(X)) \geq E^X g(X, f_o(X)). \\
 \text{Suppose for same } x', t : & \quad f_o(x) = \begin{cases} f(x) & \text{if } x \neq x' \\ t & \text{if } x = x'. \end{cases} \\
 \forall x \quad g(x, f(x)) \geq g(x, f_o(x)) & \quad \text{Choose } f \text{ to min. } g(x, f(x)). \\
 \text{Define: } f^*(x) = \arg \min_t g(x, t). & \quad \geq \underline{E^X g(X, f^*(X))}
 \end{aligned}$$

$$\begin{aligned}
 \textcircled{2} \quad \hat{Y} = f(X) \quad EL(Y, \hat{Y}) &= EL(Y, f(X)) \\
 &= \sum_{x,y} \underbrace{(y, f(x))}_{p(y|x)} \underbrace{p(x,y)}_{p(y|x)p(x)} = \sum_x \left( \sum_y (y, f(x)) p(y|x) \right) p(x) \\
 &= \sum_x g(x, f(x)) p(x) = E^X g(X, f(X)) \geq E^X g(X, f_o(X)). \\
 \text{Suppose for same } x', t : & \quad f_o(x) = \begin{cases} f(x) & \text{if } x \neq x' \\ t & \text{if } x = x'. \end{cases} \\
 \forall x \quad g(x, f(x)) > g(x, f_o(x)) & \quad \text{Choose } f \text{ to min. } g(x, f(x))
 \end{aligned}$$

$$\begin{aligned}
 & p(y|x) p(x) \overbrace{g(x, f(x))} \\
 &= \sum_x g(x, f(x)) p(x) = E^X g(X, f(X)) \geq E^X g(X, f_o(X)). \\
 \text{Suppose for same } x', t : & \quad f_o(x) = \begin{cases} f(x) & \text{if } x \neq x' \\ t & \text{if } x = x'. \end{cases}
 \end{aligned}$$

$$\forall x \quad g(x, f(x)) \geq g(x, f_o(x)) \quad \text{Choose } f \text{ to min. } g(x, f(x)).$$

Define:  $f^*(x) = \underset{t}{\operatorname{arg\min}} g(x, t).$

$\Rightarrow \boxed{p(y|x)}$  is the key quantity

$$\forall x \quad g(x, f(x)) \geq g(x, f_o(x)) \quad \text{Choose } f \text{ to min. } g(x, f(x)).$$

Define:  $f^*(x) = \underset{t}{\operatorname{arg\min}} g(x, t).$

$\Rightarrow \boxed{p(y|x)}$  is the key quantity

$$E(Y, \hat{Y}) = E^x(E(Y, \hat{Y}|X)) \leftarrow \text{L.I.E.}$$



### 3.4 Square loss

Square loss     $L(y, \hat{y}) = (y - \hat{y})^2$ .     $(X, Y) \sim P$

$$E(L(Y, \hat{y}) | X=x) = \int L(y, \hat{y}) p(y|x) dy = \int (y - \hat{y})^2 p(y|x) dy$$

$\uparrow$  smooth

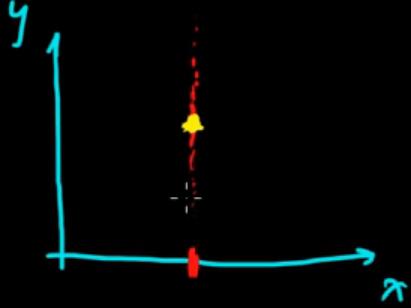
$$\begin{aligned} O &= \frac{\partial}{\partial \hat{y}} E(L(Y, \hat{y}) | X=x) = \int \frac{\partial}{\partial \hat{y}} (\hat{y} - y)^2 p(y|x) dy \\ &= \int 2(\hat{y} - y) p(y|x) dy = 2\hat{y} \underbrace{\int p(y|x) dy}_1 - 2 \underbrace{\int y p(y|x) dy}_{E(Y|X=x)} \end{aligned}$$

$$\begin{aligned} O &= \frac{\partial}{\partial \hat{y}} E(L(Y, \hat{y}) | X=x) = \int \frac{\partial}{\partial \hat{y}} (\hat{y} - y)^2 p(y|x) dy \\ &= \int 2(\hat{y} - y) p(y|x) dy = 2\hat{y} \underbrace{\int p(y|x) dy}_1 - 2 \underbrace{\int y p(y|x) dy}_{E(Y|X=x)} \\ \Rightarrow & \boxed{\hat{y} = E(Y|X=x)} \end{aligned}$$

$$\Rightarrow E(Y|X=x) = \underset{\hat{y}}{\operatorname{argmin}} E(L(Y, \hat{y}) | X=x).$$

$$\Rightarrow \hat{y} = E(Y|X=x)$$

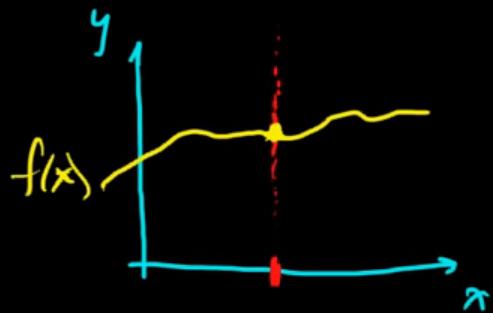
$$\Rightarrow E(Y|X=x) = \underset{\hat{y}}{\operatorname{arg\min}} E(L(Y, \hat{y}) | X=x) . \quad p(y|x)$$



$$\Rightarrow \hat{y} = E(Y|X=x)$$

$$\Rightarrow E(Y|X=x) = \underset{\hat{y}}{\operatorname{arg\min}} E(L(Y, \hat{y}) | X=x) . \quad p(y|x)$$

$$f(x) = E(Y|X=x)$$



### 3.5 The Big Picture in machine learning (part1)

- focusing on supervised learning, but same concept for others
- By starting from the decision theoretic idea that we want to minimize the expected loss between value  $y$  and our prediction,  $f(x)$ , how many of the core concepts and methods in ML fall out naturally trying to solve this problem
- key quantity:  $p(y|x)$  that we needed to solve this minimization problem

in the true  
rally while

1. Discriminative

The Big Picture  $E L(y, f(x))$   $p(y|x) \leftarrow \frac{\text{The Key}}{\text{Quantity}}$

$$D = \{(x_1, y_1), \dots, (x_n, y_n)\}$$

Discriminative Est.  $p(y|x)$  directly using  $D$ .

- kNN, Trees, SVM

Generative Est.  $p(x,y)$  using  $D$ , and then

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

Params/Latent vars  $\theta$   $P_\theta(x,y)$   $p(x,y|\theta)$

$$\underbrace{\theta, z}_{\theta}$$

Generative Est.  $p(x,y)$  using  $D$ , and then

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

Params/Latent vars  $\theta$   $P_\theta(x,y)$   $p(x,y|\theta)$

$$\underbrace{\theta, z}_{\theta}$$

$$p(y|x,D) = \int \underbrace{p(y|x,D,\theta)}_{\text{nasty}} \underbrace{p(\theta|x,D)}_{\text{nice}} d\theta$$

Often Intractable

3.6 The Big Picture (part2)

$$p(y|x,D) = \int \underbrace{p(y|x,D,\theta)}_{\text{nasty}} \underbrace{p(\theta|x,D)}_{\text{nasty}} d\theta$$

Often Intractable

↑  
positive.

↔  
nasty

↔  
nice

↔  
nasty

post. on  $\theta$



prior

## ④ Exact inference

- Multivar. Gaussian, Conjugate priors, Graphical models

## ⑤ Point est. of $\theta$

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta | x, D).$$

- MLE, MAP

$$p(y|x, D) \approx p(y|x, D, \theta_{\text{MAP}}).$$

## ⑥ Point est. of $\theta$

$$\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta | x, D).$$

- MLE, MAP

$$p(y|x, D) \approx p(y|x, D, \theta_{\text{MAP}}).$$

- Optimization, EM (Empirical Bayes)

## ⑦ Deterministic Approx.

- Laplace Approx, Variational methods, Exp. Prop.

## ⑧ Stochastic approx.

- MCMC (Gibbs, MH), Importance sampling  
(Particle filtering)

Important

### 3.7 The Big Picture (part3)

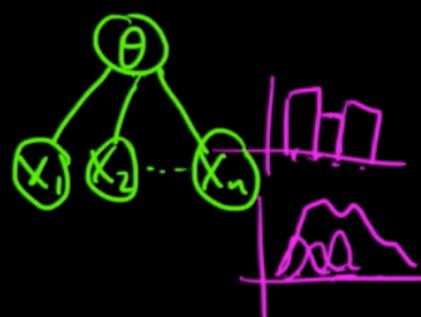
#### Density Est. (Unsupervised)

$D = (X_1, \dots, X_n)$ ,  $X_i \in \mathbb{R}^d$ , iid. Goal: Est. the distribution.

Params  $\theta$   $\rho_\theta$ .  $\theta$  random.

$$p(x|D) = \int p(x, \theta|D) d\theta$$

$$= \int p(x|\theta, D) p(\theta|D) d\theta$$



- Gen  
- Let  
- His  
- Mo  
- C

nce sampling is to estimate the expected value of the predictive p. distribution,  $p(y|x, \text{Data})$

generalize the discussion so far (which was based on the supervised setting),

's take a look at the problem of density estimation (Unsupervised)

istogram estimation, Kernel density estimation

ore prob. approach: parameters parametrizing the distribution on X's

- o Then, think of those parameters ( $\theta$ ) as random var
- o Then, suppose these data are generated from the model in this way
  - Pick a  $\theta$  from  $p_\theta$
  - Draw  $X_1, X_2, \dots, X_n$  from that sampled  $\theta$
  - Estimate the distribution over  $X$  means, to specify " $p(x|\text{all data } D)$ "

$$P(x|D) = \int P(x|\theta) P(\theta|D) d\theta$$

$$P(x|D) = \int_{\text{nasty}}^{\text{nice}} P(x|\theta) P(\theta|D) d\theta$$

posterior

$P_\theta(x) = \underbrace{P(x|\theta)}$

<--- Known

this because we choose the form of this p. distribution