Today

1. Admin, KNN
2. Probability
3. Naive Bayes

1. www.cs.ubc.ca/~schmidtm/Courses/5
2. Piazza → sign up ASAP
3. Assigment 1 DUE WEDNESDAY
   ('refresher', workload)

4. Alternative Courses

CPSC	ECE	MATH	Stat
322	592	302/303	305
340		318	447B
		418/419	460
			461
			547C

s/540

Supervised Learning

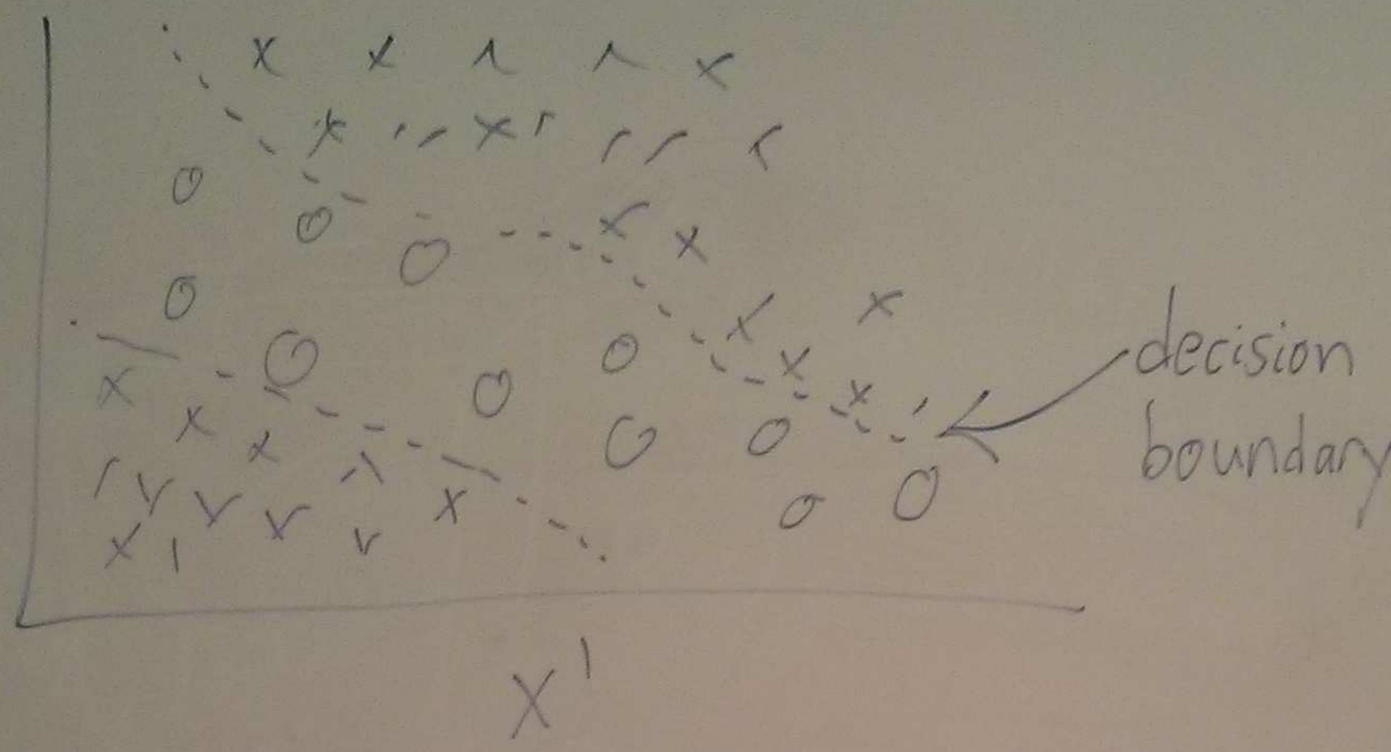
Input: X, y

Output: function $f(x_i) \mapsto y_i$

Evaluation: compare $f(x_i) = \hat{y}_i$ to y_i on <u>new</u> (x_i, y_i)

<u>Notation</u>: x_i is training example 'i' (vector)

x^i is variable 'i' (scalar)



Location based on features x^1, x^2.
+/o based on class label $y_i = +1$ or $y_i = -1$

KNN: 'closest'
- Euclidean distance $d(w,v) = \sqrt{\sum_{i=1}^{D}(w^i - v^i)^2} = \|w - v\|$
- others possible.

1-NN: 'cost'
- checking $\|x_i - x_j\|$ is $O(D)$
- do this N times for each of T test points: $O(DNT)$

k-NN: 'cost'
- sorting $O(N \log N)$
- selecting $O(kN)$

Possible course Project: Improve KNN

- better distance (learn it)
- faster classification

Q: Why can we learn?

or $y_i = -1$

$(w^i - v^i)^2 = \|w - v\|$

Today

1. Admin, kNN
2. Probability
3. Naive Bayes
4. Maximum Likelihood

Probability

$(x_i, y_i) \sim D$

$0 \leq P(A) \leq 1$

"definitely not" "definitely"

$P(\neg A) = 1 - P(A)$

We'll use $p(x) \triangleq p(X = x)$

$\sum_{x \in \mathcal{X}} p(x) = 1$

$p(A, B) = p(A \wedge B)$

$p(A \cup B) = p(A) + p(B) - p(A, B)$

Cond Prob

$$P(A|B) = \frac{P(A,B)}{P(B)} \quad \text{(for } p(B)>0\text{)}$$

'Product rule' $P(A,B) = P(A|B)P(B)$

Bayes rule $= P(B|A)P(A)$

$$P(A|B,C) = \frac{P(B|A,C)P(A|C)}{P(B|C)}$$

$$\propto P(B|A)P(A)$$

Post of A given (B and C)

$$\sum_{i=1}^{6} \frac{\prod_{j=1}^{6} \Gamma(\alpha_j + p(j))}{\Gamma(\sum_{j=1}^{6} \alpha_j)} = \frac{\prod_{j=1}^{6} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{6} \alpha_j)}$$

$p(3) = 1/6$

$p(\neg 3) = 1 - p(3) = 5/6$

$\sum_{i=1}^{6} p(i) = 1$

$p(3, \text{even}) = 0$

$p(3, \text{odd}) = 1/6$

$p(1 \cup 2) = p(1) + p(2) - p(1,2)$

$= 1/6 + 1/6 - 0$

$= 1/3$

$p(3 \mid \text{odd}) = \frac{p(3, \text{odd})}{p(\text{odd})}$

$= \frac{1/6}{1/2} = 1/3$

Marg. Prob.

$$p(A) = \sum_b p(A, B=b)$$

$$p(\text{even}) = \sum_{i=1}^{6} p(\text{even}, i)$$

$$= 0 + 1/6 + 0 + 1/6 + 0 + 1/6$$

$$= 1/2$$

We can always add extra conditioning

Conditional Independance

- $X \perp Y | Z \iff p(X, Y|Z) = p(X|Z)\, p(Y|Z)$

Today

1. Admin, KNN
2. Probability
3. Naive Bayes
4. Maximum Likelihood

Probabilistic Classifiers

Discriminative

$p(y_i | x_i)$

- no model of x
- "need lots of labelled data" but high accuracy
- e.g. logistic regression

Generative

$p(y_i, x_i)$

- need to model x
- "small amount of labelled data"
- e.g. Naive Bayes

Training

$p(y_i, x_i) = p$

Testing

$p(y_i | x_i) =$

$\propto$

Y	$p(y=0)$
0	$p(y=0)$
1	$p(y=1)$

easy

Training

$$p(x_i, y_i) = p(x_i | y_i) p(y_i)$$

$$X = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

hard

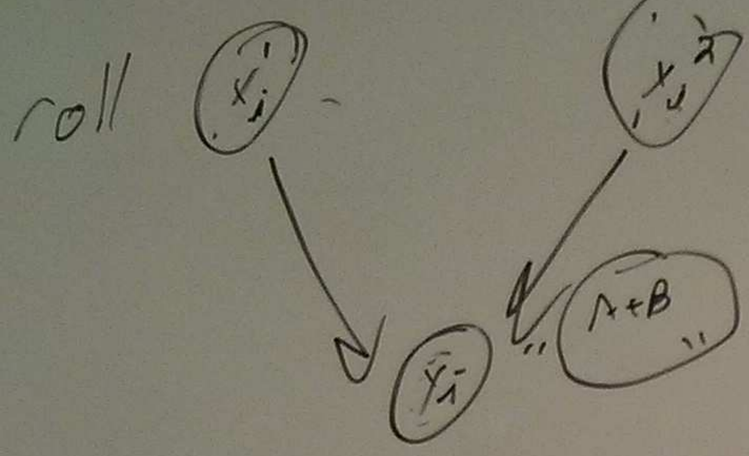
Testing

$$p(y_i | x_i) = \frac{p(x_i | y_i) p(y_i)}{p(x_i)} \propto p(x_i | y_i) p(y_i)$$

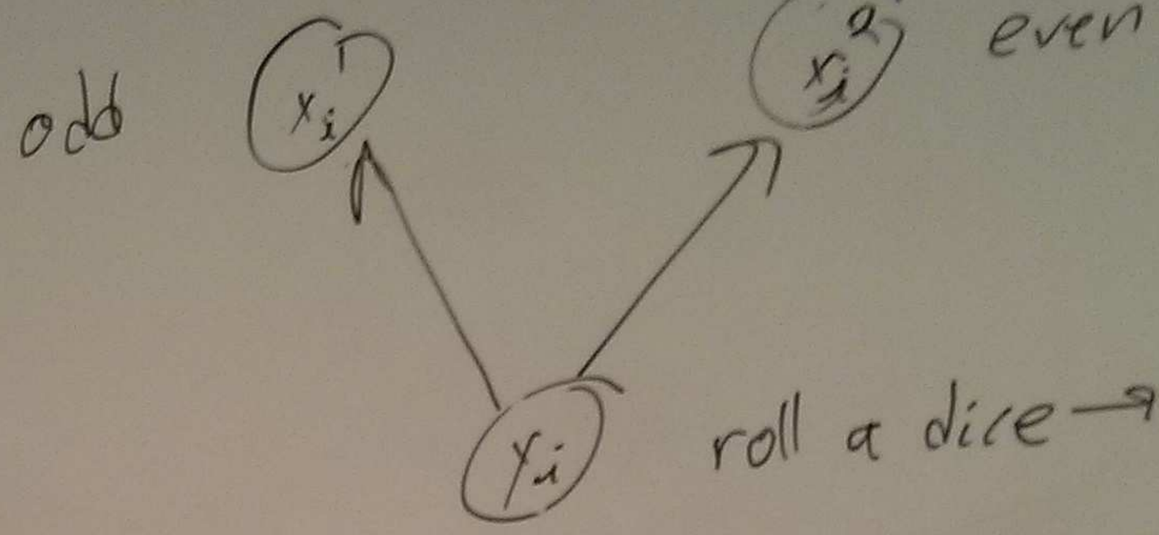
x_i^1	x_i^2	y_i	
0	0	0	$p(0,0 \mid 0)$
0	0	1	$p(0,0 \mid 1)$
0	1	0	$\vdots$
0	1	1	$\vdots$
1	0	0	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	

2^{p+1} parameters

"Discriminative"



"Generative"



roll a dice → 1, 2, ⋯, 6

Naive Bayes

x^i are "mutually independent" given y

$$\forall_{\text{subsets } A,B} \quad x^A \perp x^B \mid y$$

$$p(x_i^{1:D} \mid y_i) = p(x_i^1 \mid x_i^{2:D}, y_i)\, p(x^{2:D} \mid y_i)$$

$$= p(x_i^1 \mid y_i)\, p(x^{2:D} \mid y_i)$$

$$\vdots$$

$$= \prod_{j=1}^{D} p(x_i^j \mid y_i)$$

↳ "easy"

x^j	y
0	0
0	1
0	1
1	0
1	1

40

Maximum likelihood

$$\underset{\theta}{\operatorname{argmax}}\ p(\mathcal{D}|\theta) \overset{iid}{=} \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathcal{D}_i|\theta)$$

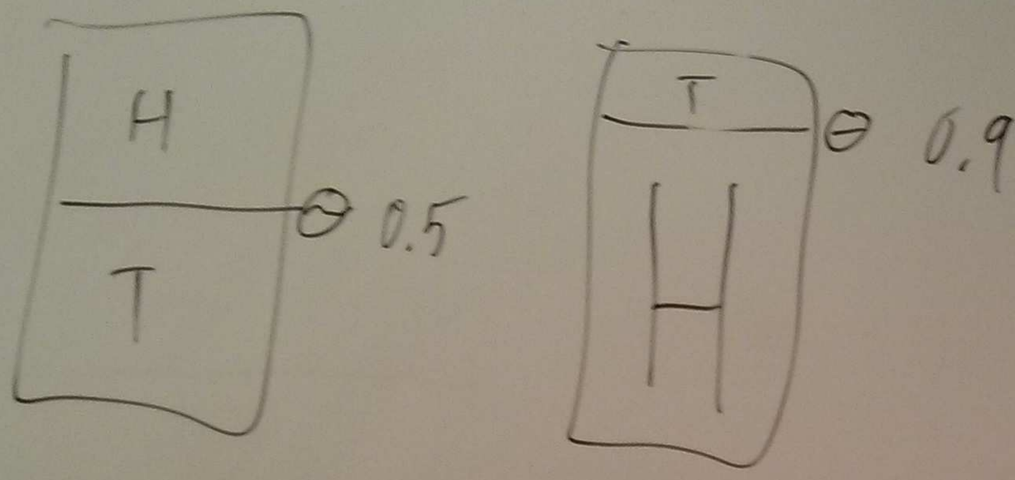
$$\Updownarrow \qquad \Updownarrow \text{ numerical}$$

$$\underset{\theta}{\operatorname{argmax}}\ \log p(\mathcal{D}|\theta) \overset{iid}{=} \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \log p(\mathcal{D}_i|\theta)$$

$$\prod_{i=1}^{N} p(x_i, y_i) \overset{}{=} \prod_{i=1}^{N} p(x_i|y_i)p(y_i)$$

$$= \prod_{i=1}^{N} p(y_i) \prod_{j=1}^{p} p(x_i^j|y_i)$$

$$= \prod_{i=1}^{N} Ber(y_i|\theta) \prod_{j=1}^{p} Ber(x_i^j|y_i, \theta_j^{y_i})$$



$$\log p(X, Y) = \sum_{i=1}^{N} \log p(Y_i \mid \theta) + \sum_{j=1}^{D} \log p(X_i^j \mid Y_i, \theta_j^{Y_i})$$

θ: p("head")

$$Y_i \mid \theta \sim Ber(\theta)$$

$$p(Y_i \mid \theta) = \theta^{\overbrace{I(Y_i=1)}^{\text{"head"}}} (1-\theta)^{\overbrace{I(Y_i=0)}^{\text{"tails"}}}, \quad 0 \le \theta \le$$

$$N_1 + N_0 = N$$

$$(1-\theta) N_1 = N_0 \theta$$

$$\frac{N_1/N}{N_0/N} = \frac{\theta}{1-\theta}$$

$$\theta = \frac{N_1}{N}$$

$$y_i \in \{1, 2, 3, 4\} \qquad \begin{matrix} c=1 & c=2 & \cdots \end{matrix}$$

$$x_i | y_i, \theta_j^{y_i}$$

$$\theta^c = \frac{N_c}{N} \qquad \theta_j^c = \frac{N_{jc}}{N_c}$$

$2(1-\theta)]$