1. What is optimization? What is a "loss function"?

optimization in ML: finding the optimal parameters of the model that give us the most accurate results (predictions)

loss function: we define a function that best describe our undesirable outcomes (errors) for each model.
(cost function)

2. Given the following dataset, build a Naïve Bayes model for the given training instances.

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
A	s	h	n	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
G	o	m	n	T	?
H	?	h	?	F	?

6 training instances

NB : probabilistic classification model

$\Rightarrow$ Calculate prior & conditional probs

Prior probs:

$P(Y) = \frac{1}{2}$ $\quad$ $P(N) = \frac{1}{2}$

Conditional:

$P(Outl=s \mid Y) = 0$ $\quad$ $P(Outl=s \mid N) = \frac{2}{3}$

$P(Outl=o \mid Y) = \frac{1}{3}$ $\quad$ $P(Outl=o \mid N) = 0$

$P(Outl=r \mid Y) = \frac{2}{3}$ $\quad$ $P(Outl=r \mid N) = \frac{1}{3}$

$P(\text{Temp} = h \mid N) = \frac{2}{3}$ $P(\text{Temp} = m \mid N) = 0$ $P(\text{Temp} = c \mid N) = \frac{1}{3}$

$P(\text{Temp} = h \mid Y) = \frac{1}{3}$ $P(\text{Temp} = m \mid Y) = \frac{1}{3}$ $P(\text{Temp} = c \mid Y) = \frac{1}{3}$

$P(\text{Humi} = n \mid N) = \frac{2}{3}$ $P(\text{Humi} = h \mid N) = \frac{1}{3}$

$P(\text{Humi} = n \mid Y) = \frac{1}{3}$ $P(\text{Humi} = h \mid Y) = \frac{2}{3}$

$P(\text{Wind} = T \mid N) = \frac{2}{3}$ $P(\text{Wind} = F \mid N) = \frac{1}{3}$

$P(\text{Wind} = T \mid Y) = 0$ $P(\text{Wind} = F \mid Y) = 1$

3. Using the Naïve Bayes model that you developed in question 2, classify the given test instances.

(i). No smoothing.

$$\hat{c} = \underset{c}{\operatorname{argmax}}\ P(c \mid \underbrace{t_1, \dots, t_k}_{\text{attributes}})$$

$$= \underset{c}{\operatorname{argmax}}\ \frac{P(t_1, \dots, t_k \mid c)\, P(c)}{P(t_1, \dots, t_k)}$$

same for all classes (ignore in "argmax")

$$= \underset{c}{\operatorname{argmax}}\ P(t_1, \dots, t_k \mid c)\, P(c)$$

$$= \underset{c}{\operatorname{argmax}}\ \underbrace{\left\{ \prod_{i=1}^{k} P(t_i \mid c) \right\}}_{\substack{\text{NB assumption:} \\ \text{attributes are conditionally independent}}} \cdot P(c)$$

Test Instances					
G	o	m	n	T	?
H	?	h	?	F	?

$P(\text{Outl} = s \mid N) = \frac{2}{3}$ $P(\text{Outl} = o \mid N) = 0$ $P(\text{Outl} = r \mid N) = \frac{1}{3}$

$P(\text{Outl} = s \mid Y) = 0$ $P(\text{Outl} = o \mid Y) = \frac{1}{3}$ $P(\text{Outl} = r \mid Y) = \frac{2}{3}$

$P(\text{Temp} = h \mid N) = \frac{2}{3}$ $P(\text{Temp} = m \mid N) = 0$ $P(\text{Temp} = c \mid N) = \frac{1}{3}$

$P(\text{Temp} = h \mid Y) = \frac{1}{3}$ $P(\text{Temp} = m \mid Y) = \frac{1}{3}$ $P(\text{Temp} = c \mid Y) = \frac{1}{3}$

$P(\text{Humi} = n \mid N) = \frac{2}{3}$ $P(\text{Humi} = h \mid N) = \frac{1}{3}$

$P(\text{Humi} = n \mid Y) = \frac{1}{3}$ $P(\text{Humi} = h \mid Y) = \frac{2}{3}$

$P(\text{Wind} = T \mid N) = \frac{2}{3}$ $P(\text{Wind} = F \mid N) = \frac{1}{3}$

$P(\text{Wind} = T \mid Y) = 0$ $P(\text{Wind} = F \mid Y) = 1$

$P(Y) = \frac{1}{2}$

$P(N) = \frac{1}{2}$

Instance G:

- Class Y:

$P(Outl=o|Y)\,P(Temp=m|Y)\,P(Humi=n|Y)$

$\times P(Wind=T|Y)\,P(Y)$

$= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times 0 \times \frac{1}{2} = 0$

- Class N:

$P(\cdot\cdot) \cdots P(N)$

$= 0 \times 0 \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{2}$

$= 0$

$\Rightarrow$ Choose either label for G

(e.g. Y)

Instance H:

missing values $\Rightarrow$ use what we have.

- Class Y:

$P(Temp=h|Y)\,P(wind=F|Y)\,P(Y)$

$= \frac{1}{3} \times 1 \times \frac{1}{2} = \frac{1}{6}$ ✓

- Class N:

$P(Temp=h|N)\,P(wind=F|N)\,P(N)$

$= \frac{2}{3} \times \frac{1}{3} \times \frac{1}{2} = \frac{1}{9}$

$\Rightarrow$ Classify instance H as Y.

(ε)

(ii). Using the "epsilon" smoothing method.

Replace zero probabilities with ε (a small value, e.g. 10^{-6})

For instance G:

Y: $P(Outl=o|Y)\,P(Temp=m|Y)\,P(Humi=n|Y)$

$\times P(Wind=T|Y)\,P(Y)$

$= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \varepsilon \times \frac{1}{2} = \frac{\varepsilon}{54}$

N: $P(\cdot\cdot) \cdots P(N)$

$= \varepsilon \times \varepsilon \times \frac{2}{3} \times \frac{2}{3} \times \frac{1}{2}$

$= \frac{2\varepsilon^2}{9}$

$\frac{\varepsilon}{54} > \frac{2\varepsilon^2}{9}$ $\Rightarrow$ Classify G as Y.

For instance H: no zero probabilities $\Rightarrow$ still Y.

add 1 smoothing

(iii). Using "Laplace" smoothing ($\alpha = 1$)

For conditional probs:

$$p_i = \frac{x_i + \alpha}{N + \alpha d} \quad (\text{was } p_i = \frac{x_i}{N})$$

d ← # possible values for this attribute

$\Rightarrow$ ensure $\Sigma_K P(A_i = K \mid \text{Class } c) = 1$

$\Rightarrow$ No zero conditional probs! (If $x_i = 0$, $p_i = \frac{\alpha}{N + \alpha d}$)

Instance G:

s, o, r — h, m, c — n, h

Y: $P(\text{Outl} = o \mid Y)\, P(\text{Temp} = m \mid Y)\, P(\text{Humi} = n)$

$\times P(\text{Wind} = T)\, P(Y)$

T, F

$$= \frac{1+1}{3+3} \times \frac{1+1}{3+3} \times \frac{1+1}{3+2} \times \frac{0+1}{3+2} \times \frac{1}{2}$$

$$= \frac{2}{6} \times \frac{2}{6} \times \frac{2}{5} \times \frac{1}{5} \times \frac{1}{2}$$

$= 0.0044$

N: $P(\cdot\cdot) \cdots P(N)$

$$= \frac{0+1}{3+3} \times \frac{0+1}{3+3} \times \frac{2+1}{3+2} \times \frac{2+1}{3+2} \times \frac{1}{2}$$

$$= \frac{1}{6} \times \frac{1}{6} \times \frac{3}{5} \times \frac{3}{5} \times \frac{1}{2}$$

$= 0.005$ ✓

$\Rightarrow$ Classify instance G as N.

Similarly for H ...

4. For the following set of classification problems, we want to design a Naive Bayes classification model.

(iv). You want to classify a set of images of animals in to 'cats', 'dogs', and 'others'.

(v). You want to classify whether each customer will purchase a product, given all the products (s)he has bought previously.

Answer the following questions for each problem:

(1) what are the instances, what are the features (and values)?

(2) explain which distributions you would choose to model the observations, and

(3) explain the significance of the Naive Bayes assumption.

(1) Instances : images

Features : pixels of images

values : intensity color shade

continuous

(2) Gaussian distribution for each feature value : intensity, color, ...

(continuous feature, assume feature values are normally distributed)

(3) NB assumption : Conditional independence assumption (among features)

$$P(\underbrace{t_1, \dots, t_k}_{\text{features}} \mid c) = \prod_{i=1}^{k} (t_k \mid c)$$

class (e.g. cat)

⇒ Not true in the reality : intesity, color, ..., of neighboring pixels depend on one another

⇒ Not independent, but we can still use NB to make predictions (v). You want to classify whether each customer will purchase a product, given all the products (s)he has bought previously.

(1) Instances : Customer

Features : Products (binary , or # of times the customer bought)

(2) K - dim multinomial distribution : values are the counts of purchases of each product

(or bernoulli dist for each variable)

(3) Under NB assumption : the purchases of products are independent

Not independent ⇒ e.g. (Milk & bread)