Naïve Bayes

ECE/CS 498 DS U/G

Lecture 5

Professor Ravi K. Iyer

University of Illinois

Announcements

- MP1 Task 0 and 1 were due yesterday (2/3). Hope it went well!

- MP1 Checkpoint 2 is due on Feb 10

- Discussion section on Friday, Feb 8
  - Students attending previous sessions found it helpful; everyone is encouraged to attend

- In Class Activity 1 submissions due today
  - Please submit them in the box at the back of the lecture auditorium

- HW0 grades released on Compass2G

- Turn in no later than one week for HW regrade

Mean	Std	Median	Max	Min
93.5	10.6	96.0	100	25
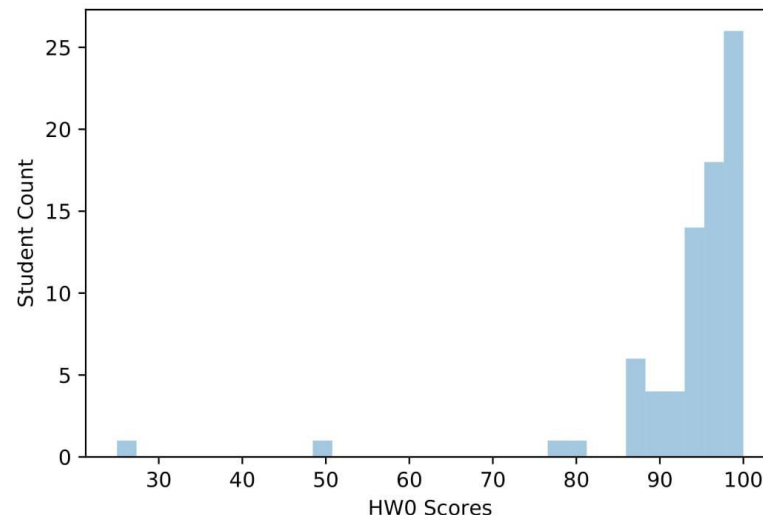
Naïve Bayes- Intuitive Example: Infer the Dog-Breed

- Assume that we have two classes (breeds): C_1 = German Shepherd (GS), and C_2 = Dalmatian (D)

- We define *tall* dogs as having a height more than 40 cm

- We have a dog whose breed we do not know, say d

- We know that d is ***tall***

- Based on the above information and *training data*, can we infer the breed of d?

- **Equivalent Question**: Based on the training data, Is it more probable that the ***tall*** dog is a German Shepherd or a Dalmatian, i.e., which is greater $p(GS|tall)$ or $p(D|tall)$

German Shepherd

Dalmatian

Think about Bayes Theorem

What is the probability of being tall given Dalmatian?

What is the probability of being a Dalmatian?

$$p(D|tall) = \frac{p(tall|D)p(D)}{p(tall)}$$

What is the probability of being tall?

- Fortunately, we have a database with dog heights and breed

- We use the database to calculate the posterior probability for both D and GS

Height	Breed
Short	GS
Tall	D
Short	D
Tall	D
Tall	GS
Short	D
Short	D
Short	GS

Guess the dog breed: the calculations

$$p(C_k|tall) = \frac{p(tall|C_k)\,p(C_k)}{p(tall)}.$$

C_k=Class k

$$p(GS|tall) = \frac{p(tall|GS)\,p(GS)}{p(tall)} = \frac{\frac{1}{3} * \frac{3}{8}}{3/8} = \frac{0.125}{3/8}$$

$$p(D|tall) = \frac{p(tall|D)\,p(D)}{p(tall)} = \frac{\frac{2}{5} * \frac{5}{8}}{3/8} = \frac{0.250}{3/8}$$

Height	Breed
Short	GS
Tall	D
Short	D
Tall	D
Tall	GS
Short	D
Short	D
Short	GS

$p(D|tall) > p(GS|tall)$, therefore, dog d is more likely to be Dalmatian.

Multiple features

- In our (2 – breed) example we had just one *feature, i.e., height*
- What if we have several *features e.g., height, weight, color of eyes*?
- Say that the dog d is *tall, heavy* and has *brown eyes*
- Lets assume that **features are independent,** given the breed, we can write

$$p(d|C_k) = p(tall, heavy, browneyes|C_k)$$
$$= p(tall|C_k)p(heavy|C_k)p(browneyes|C_k)$$

- More generally (with n features),

$$p(\boldsymbol{x}|C_k) = p(x_1|C_k) * p(x_2|C_k) * \cdots * p(x_n|C_k)$$

Probability of class C_k generating instance x

Probability of class C_k generating observed value of feature 1

Probability of class C_k generating observed value of feature 2

Naïve Bayes Classifier Representation

- The assumption of independence of features given the class results in a **Naïve Bayes classifier**
- Naïve Bayes classifier can be represented as a graph
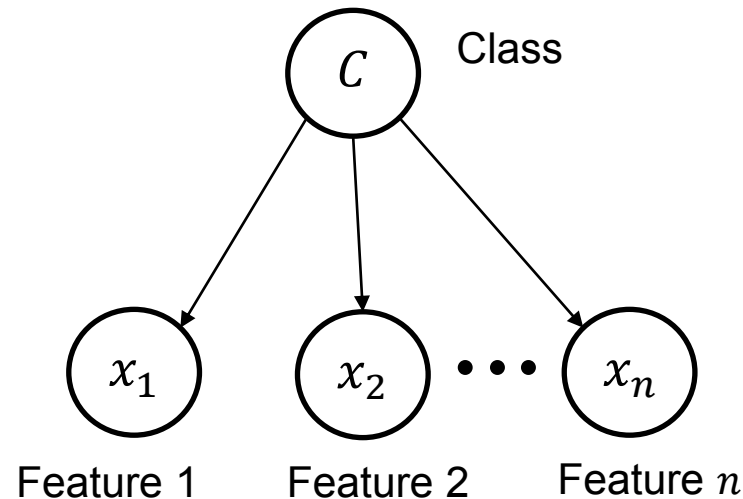
Class

Note the direction of the arrow

Feature 1   Feature 2   Feature n

Naïve Bayes Classifier (Cont'd)

- We want to answer the question - Given a **previously unseen** data point, which class does the data point belong to?

- Previously unseen data point: $x = (x_1, x_2, \ldots, x_n)$
  K classes: $C_k, 1 \leq k \leq K$

- Intuitively, the most probable class would be the one with the maximum probability given the observations (data or evidence)

$$C^* = \underset{k \in \{1,..,K\}}{\mathrm{argmax}} \; p(C_k | x)$$

- There are three main concepts here in deriving the Naïve Bayes classifier
  - Chain rule
  - Conditional Independence
  - Maximum a posteriori (MAP) rule

Bayes Theorem and Chain Rule

From Bayes Theorem, we have

$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}$$

The above involves calculation of $p(\boldsymbol{x}|C_k) = p(x_1, x_2, \ldots, x_n|C_k)$

Using **chain rule**, we get

$$
\begin{aligned}
p(\boldsymbol{x}|C_k) &= p(x_1, x_2, \ldots, x_n|C_k) \\
&= p(x_1|x_2, \ldots, x_n, C_k)p(x_2, \ldots, x_n|C_k) \\
&= p(x_1|x_2, \ldots, x_n, C_k)p(x_2|x_3, \ldots, x_n, C_k)p(x_3, \ldots, x_n|C_k)\ldots \\
&= \ldots \\
&= p(x_1|x_2, \ldots, x_n, C_k)p(x_2|x_3, \ldots, x_n, C_k)\ldots p(x_{n-1}|x_n, C_k)p(x_n|C_k)
\end{aligned}
$$

Conditional Independence

- NB assumes that the features are class conditionally independent, i.e., independent given the class
- x_i is conditionally independent of x_j given C_k if $i \neq j$, $\forall\, i, j \in \{1, \dots, n\}$
- Therefore, $p\left(x_i \middle| x_j, C_k\right) = p(x_i | C_k)$ if $i \neq j$
- Therefore, $p(x_i | x_{i+1}, x_{i+2}, \dots, x_n, C_k) = p(x_i | C_k), \forall\, i \in \{1, \dots, n-1\}$
- Applying the class conditional independence to $p(\boldsymbol{x} | C_k)$ gives

$$p(\boldsymbol{x}|C_k) = p(x_1|x_2, \dots, x_n, C_k)p(x_2|x_3, \dots, x_n, C_k) \dots p(x_{n-1}|x_n, C_k)p(x_n|C_k)$$

$$= p(x_1|C_k)p(x_2|C_k) \dots p(x_n|C_k)$$

$$= \prod_{i=1}^{n} p(x_i|C_k)$$

MAP

- We started with: $p(C_k|\boldsymbol{x}) = \dfrac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}$

- Let $Z = p(\boldsymbol{x})$. We get,

$$p(C_k|\boldsymbol{x}) = \frac{1}{Z}p(C_k)\prod_{i=1}^{n}p(x_i|C_k)$$

- Apply the MAP rule, i.e., pick the class with maximum posterior probability

$$C^* = \underset{k\in\{1,..,K\}}{\mathrm{argmax}}\, p(C_k|\boldsymbol{x}) = \underset{k\in\{1,..,K\}}{\mathrm{argmax}}\frac{1}{Z}p(C_k)\prod_{i=1}^{n}p(x_i|C_k)$$
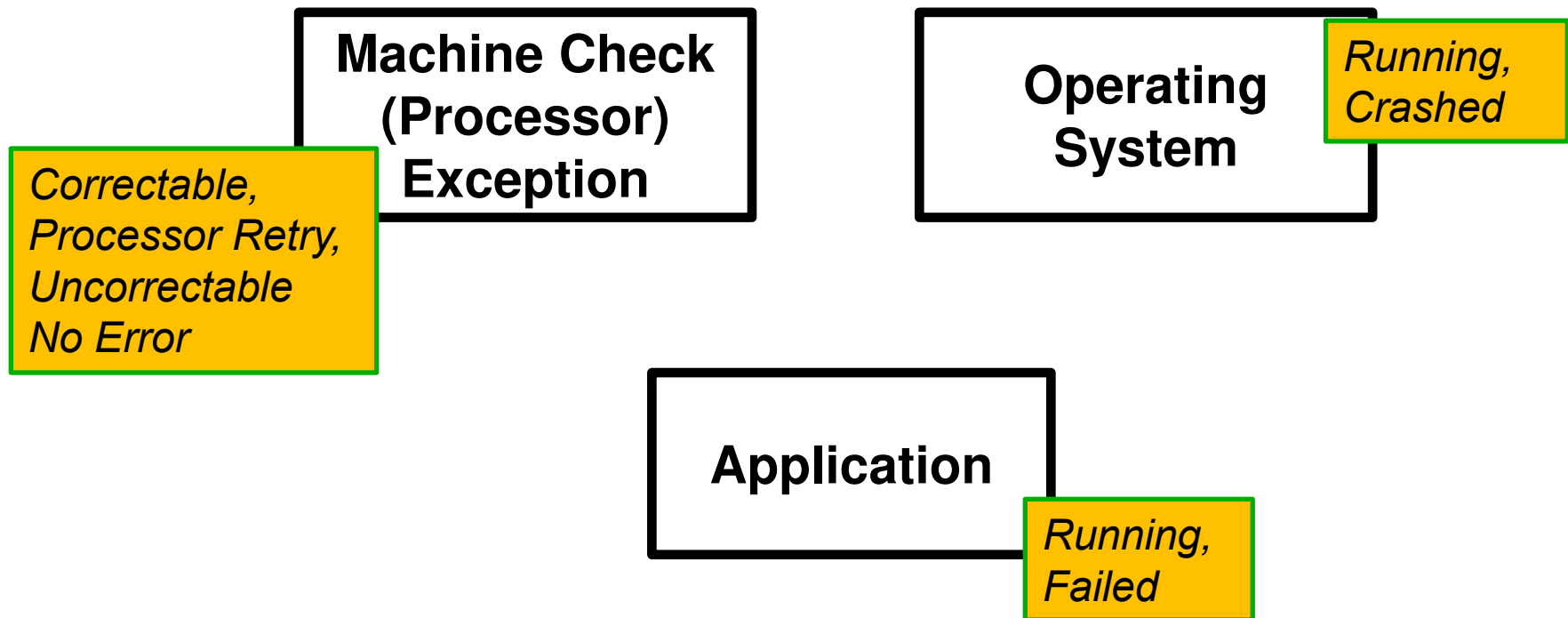
MAP

- We have computed everything except Z
  - $p(x_i|C_k)$ and $p(C_k)$ are computed from the training data

- Z is dependent only on x (no dependence on k). Therefore, it is constant once x is known and can be ignored.

- Naïve Bayes Classifier is:

$$C^* = \underset{k \in \{1,\dots,K\}}{\mathrm{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

Naïve Bayes Model Depicting the Resilience of your Laptop Application

- *Task: Predict application exit status (success, failure) based on the observed failures in the system (such as MCEs, and OS errors)*

Machine Check (Processor) Exception

Operating System

Running, Crashed

Correctable, Processor Retry, Uncorrectable No Error

Application

Running, Failed

Example of Machine Check Exception (Processor Error)

A problem has been detected and windows has been shut down to prevent damage to your computer.

MACHINE_CHECK_EXCEPTION

If this is the first time you've seen this stop error screen, restart your computer. If this screen appears again, follow these steps:

Check to make sure any new hardware or software is properly installed. If this is a new installation, ask your hardware or software manufacturer for any windows updates you might need.

If problems continue, disable or remove any newly installed hardware or software. Disable BIOS memory options such as caching or shadowing. If you need to use Safe Mode to remove or disable components, restart your computer, press F8 to select Advanced Startup Options, and then select Safe Mode.

Technical information:

*** STOP: 0x0000009C (0x0000000000000000,0xFFFFFADF90A81240,0x00000000B2000040,0x0000000000000800)
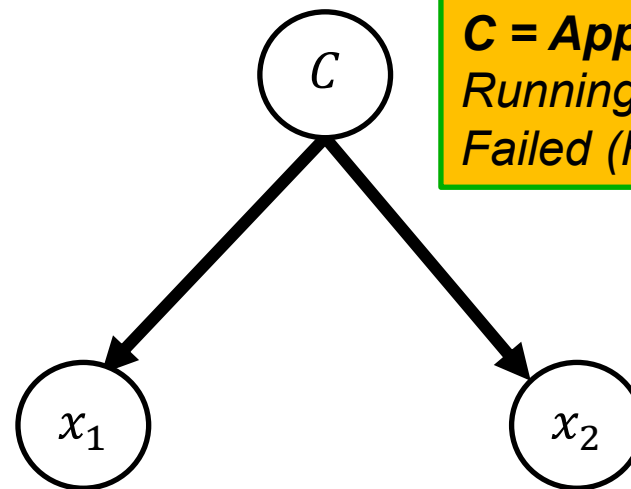
Beginning dump of physical memory

Physical memory dump complete.
Contact your system administrator or technical support group for further assistance.

Naïve Bayes

- Predicting application failures given MCE and OS-information

C = **Application Status**
Running (R),
Failed (F)

x_1**: Machine Check Exception**
Correctable (C), Processor Retry (P),
Uncorrectable (U), No Error (NE)

x_2**: Operating System Status**
Running (R),
Failed (F)

Toy ML for Personal Computers: Data

MacOS: sudo cat /var/log/system.log
Linux: sudo cat /var/log/syslog

1363323625 local3 6 2013-03-15T00:00:25.07 c6-3c1s1n0 xtconsole 11144 p0-20130219t183043 CPU 16: **Machine Check Exception: 0 Bank 4: 9c6244000Ic017b**

1363323625 local3 6 2013-03-15T00:00:25.07 c6-3c1s1n0 xtconsole 11144 p0-20130219t183043 TSC 0 ADDR bcb75daa0 MISC c00a000001000000

1363323625 local3 6 2013-03-15T00:00:25.07 c6-3c1s1n0 xtconsole 11144 p0-20130219t183043 **PROCESSOR** 2:600f12 TIME 1363323624 SOCKET 1 APIC 20
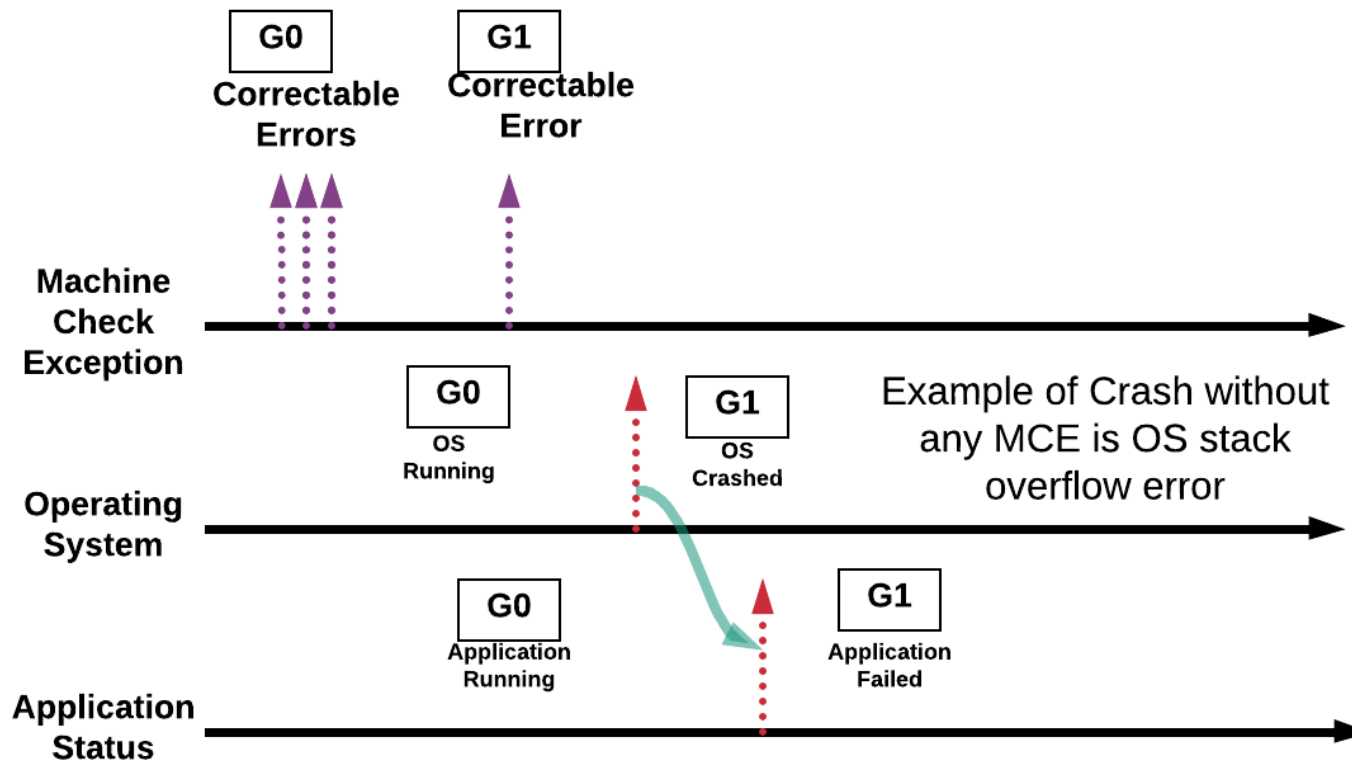
1452290276 daemon 5 2016-01-08T15:57:56-06:00 172.16.255.118 puppet-agent 25348 - (/Stage[main]/Sysctl/Exec[subscribe-sysctl]/returns) **kernel.panic = 0**

1452290276 local3 6 2015-09-25T00:00:37.369592-05:00 c8-6c2s7n3   **APP=NAMD, exit_code = 1**

Coalescing Like Events
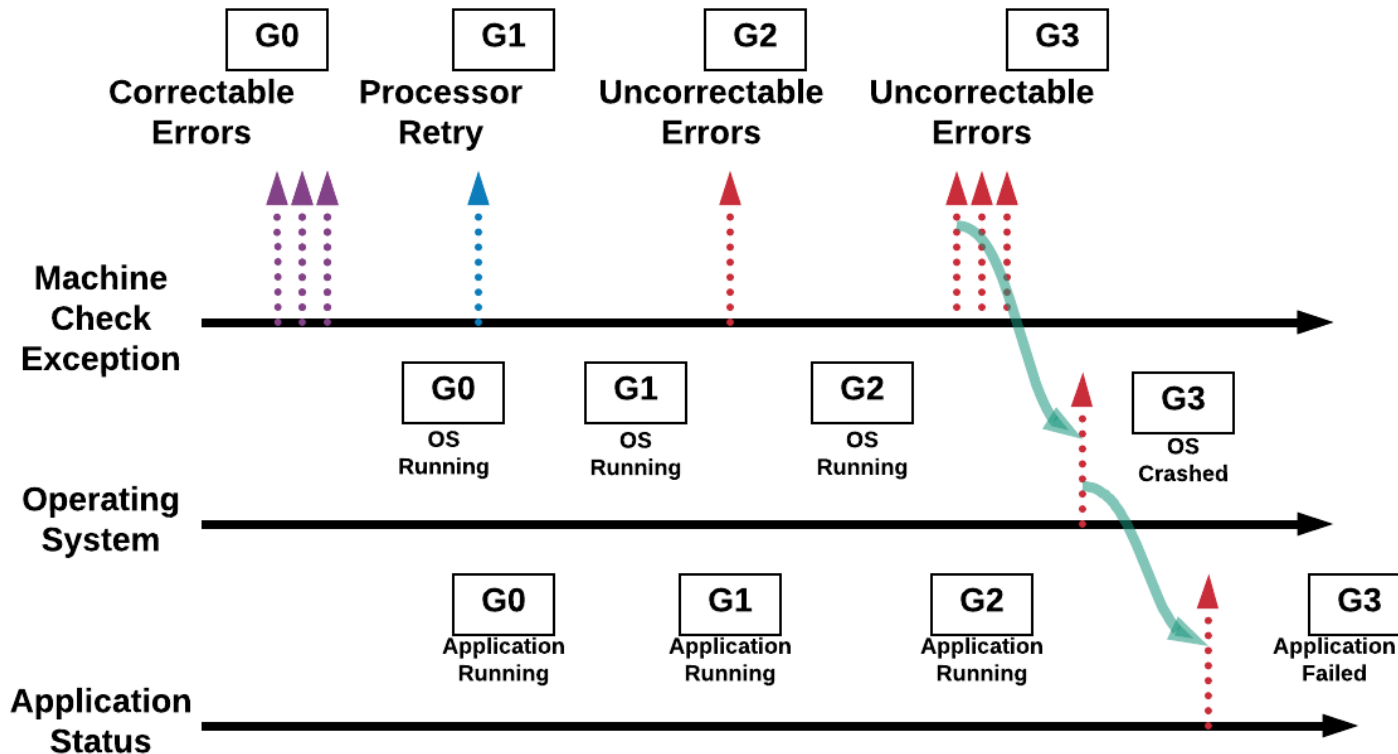
Coalescing Like Events

Final Dataset

Cluster	MCE	OS	App
G0	C	R	**R**
G1	P	R	**R**
G2	U	R	**R**
G3	U	F	**R**
…	..	..	..
G999	C	F	**F**

Naïve Bayes Training

- Recall Naive Bayes Classifier

$$C^* = \underset{k \in \{1,\ldots,K\}}{\mathrm{argmax}}\, p(C_k) \prod_{i=1}^{n} p(x_i | C_k)$$

- In our example, we need to train and get following parameters

  – Prior probabilities : $P(App = Running), P(App = Fail)$

  – Conditional Probabilities for each feature

$P(OS = Running|App = F)$	$P(OS = Running|App = R)$
$P(OS = Failed|App = F)$	$P(OS = Failed|App = R)$

$P(MCE = Correctable|App = F)$	$P(MCE = Correctable|App = R)$
$P(MCE = Processor\ Retry|App = F)$	$P(MCE = Processor\ Retry|App = R)$
$P(MCE = Uncorrectable|App = F)$	$P(MCE = Uncorrectable|App = R)$
$P(MCE = No\ error|App = F)$	$P(MCE = No\ error|App = R)$

Training the Model

$$C = \underset{k \in \{1,..,K\}}{\mathrm{argmax}} \; P(C_k) \prod_{i=1}^{n} P(x_i | C_k)$$

Cluster	MCE	OS	App
G0	C	R	**R**
G1	P	R	**R**
G2	U	R	**R**
G3	U	F	**R**
…	..	..	..
G999	C	F	**F**

Counting

Prior Probabilities
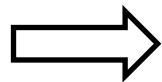
App	#	P(App)
R	950	0.95
F	50	0.05

CPTs

MCE	App	#	P(MCE|App)
C	R	66	0.07
P	R	5	0.005
U	R	5	0.005
NE	R	874	0.92
C	F	2	0.04
P	F	5	0.1
U	F	40	0.8
NE	F	3	0.06
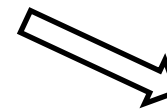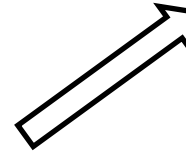
OS	App	#	P(OS|App)
R	R	940	0.99
F	R	10	0.01
R	F	45	0.9
F	F	5	0.1

Naïve Bayes Inference – Example 1

Inference task: Is application running or failed ?

C1: $P(MCE|App = F)\ P(OS|App = F)\ P(App = F)$
C2: $P(MCE|App = R)P(OS|App = R)\ P(App = R)$

- Given instance:
    - MCE = P, OS = R
- Posterior probabilities
    - $P(C_1|\boldsymbol{x}) \propto P(MCE = P|App = \boldsymbol{F})\ P(OS = R|App = \boldsymbol{F})\ P(App = \boldsymbol{F})$
    - $P(C_2|\boldsymbol{x}) \propto P(MCE = P|App = \boldsymbol{R})\ P(OS = R|App = \boldsymbol{R})\ P(App = \boldsymbol{R})$
- Evaluation via substitution
    - $P(C_1|\boldsymbol{x}) \propto$ 0.1 * 0.9 * 0.05 = 0.0045
    - $P(C_2|\boldsymbol{x}) \propto$ 0.005 * 0.99 * 0.95 = **0.0047025**
- Apple MAP Rule: Select hypothesis corresponding to max P
    - **C2 selected, i.e., APP = R (running)**

Naïve Bayes Inference – Example 2

Inference task: Is application running or failed ?

C1: $P(MCE|App = F)\ P(OS|App = F)\ P(App = F)$
C2: $P(MCE|App = R)P(OS|App = R)\ P(App = R)$

- Given instance:
  - MCE = U, OS = F
- Posterior probabilities
  - $P(C_1|\boldsymbol{x}) \propto P(MCE = U|App = \textbf{\textit{F}})\ P(OS = F|App = \textbf{\textit{F}})\ P(App = \textbf{\textit{F}})$
  - $P(C_2|\boldsymbol{x}) \propto P(MCE = U|App = \textbf{\textit{R}}) \times P(OS = \text{F}|App = \textbf{\textit{R}})\ P(App = \textbf{\textit{R}})$
- Evaluation via substitution
  - $P(C_1|\boldsymbol{x}) \propto$ 0.8 * 0.1 * 0.05 = **4.0e-03**
  - $P(C_2|\boldsymbol{x}) \propto$ 0.005 * 0.01 * 0.95 = 4.75e-05
- Apple MAP Rule: Select hypothesis corresponding to max P
  - **C1 selected, i.e., APP = F (failed)**

Back to the Dogs

- We have two classes: C_1 = German Shepherd (GS), and C_2 = Dalmatian (D)

- Instead of categorizing height as "tall" or "short", we use its exact value

- We have a dog whose breed we do not know, say $\tilde{d}$

- We know that $\tilde{d}$'s height is 38 cm

- Based on the above information and some training data, can we find the breed of $\tilde{d}$?
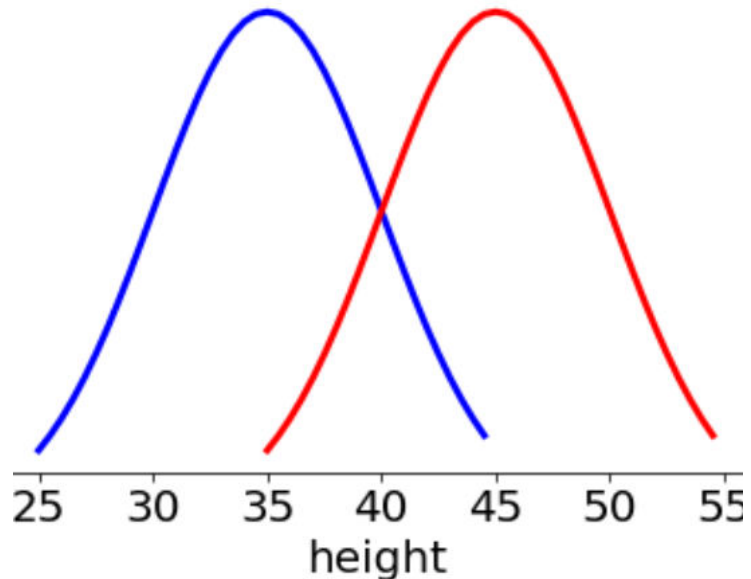

German Shepherd

Dalmatian

Naïve Bayes with continuous variables

- Height can be any real number; therefore we its **distribution** will be **continuous**

- With enough data, approximate the distribution of height for the two classes

  - We can also assume a parametric form for the distribution

- Following is the distribution of height based on some training data

  - German Shepherd: $\mathcal{N}(\mathbf{35}, \mathbf{25})$
  - Dalmatian: $\mathcal{N}(\mathbf{45}, \mathbf{25})$

- Priors are equal

$$p(C_k|38) = \frac{p(38|C_k)\,p(C_k)}{p(38)}$$

$$p(GS|38) = \frac{p(38|GS)\,p(GS)}{p(38)}$$

$$\propto \frac{1}{\sqrt{50\pi}}\exp\left(-\frac{(38-35)^2}{50}\right)$$

$$*\,0.5 = 0.067 * 0.5 = \mathbf{0.034}$$

$$p(D|38) = \frac{p(38|D)\,p(D)}{p(38)}$$

$$\propto \frac{1}{\sqrt{50\pi}}\exp\left(-\frac{(38-45)^2}{50}\right)$$

$$*\,0.5 = 0.03 * 0.5 = 0.015$$

$p(D|38) < p(GS|38)$, therefore, dog $\tilde{d}$ is more likely to be a German Shepherd.