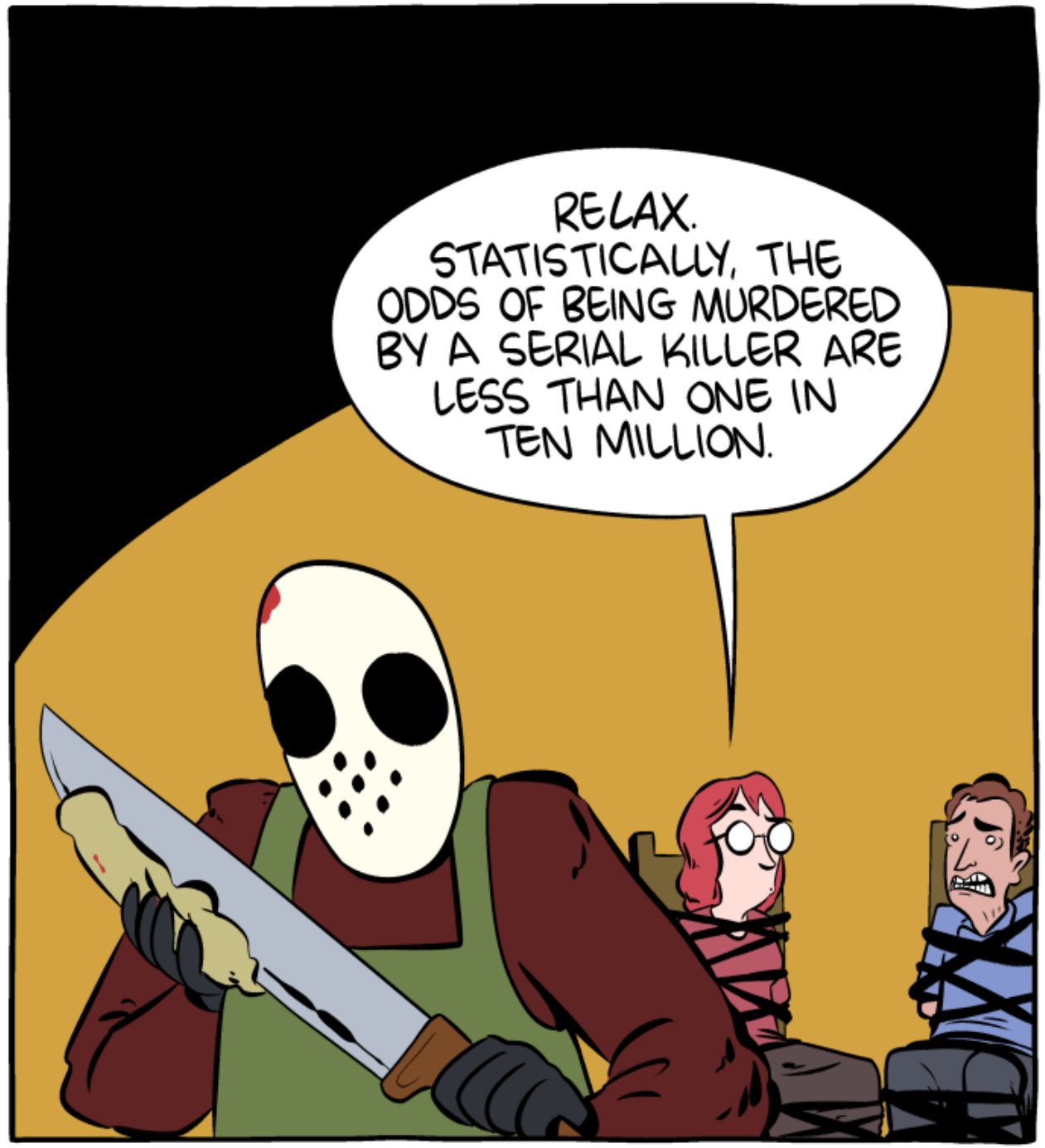
Naïve Bayes Classification

Understanding the Basics with an Illustrative Example (Breast Cancer Dataset)

Shintaro Hagiwara, Patrick Boily



Illustrating NBC With an Example

Preliminaries – Breast Cancer Dataset

2 types of tumors: **Benign** and **Malignant**.

9 variables are measured: **clump thickness**, **uniformity of cell size**, **mitoses**, etc.

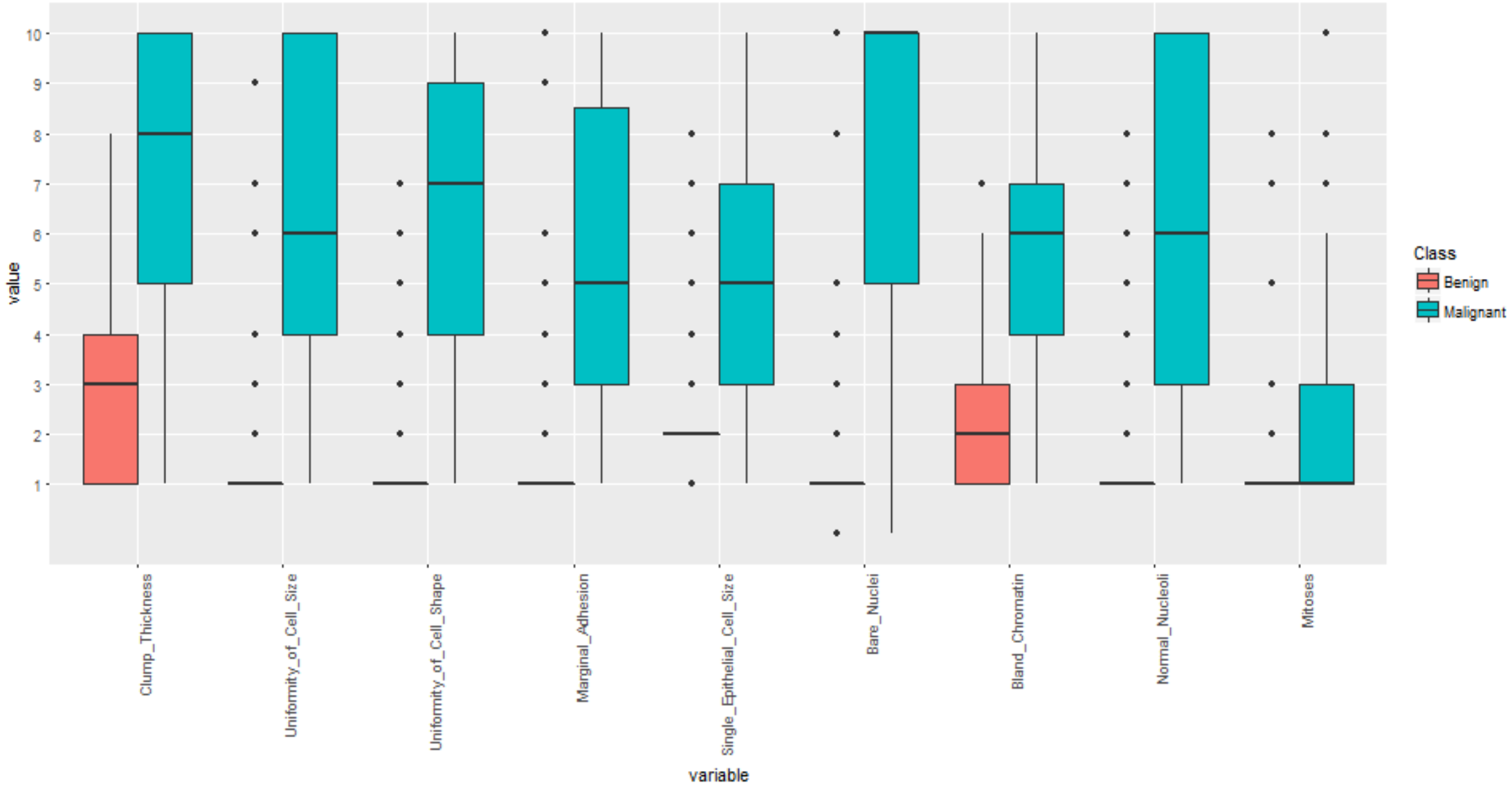
All measurements are given in scale of 1 to 10 (missing = 0).

699 samples are collected, of which 458 are benign, and 241 are malignant.

559 cases are used to build the classifier, and tested using 140 data points.

Illustrating NBC With an Example

Preliminaries – Visualizing the Data



Illustrating NBC With an Example
Preliminaries – Objective

Question: we have 140 undiagnosed subjects with measurements. Can we determine whether their tumor is benign or malignant based on the score?

Advantages:

- easy to implement
- naïve (simple assumptions)
- robust
- … and in some sense, **optimal**

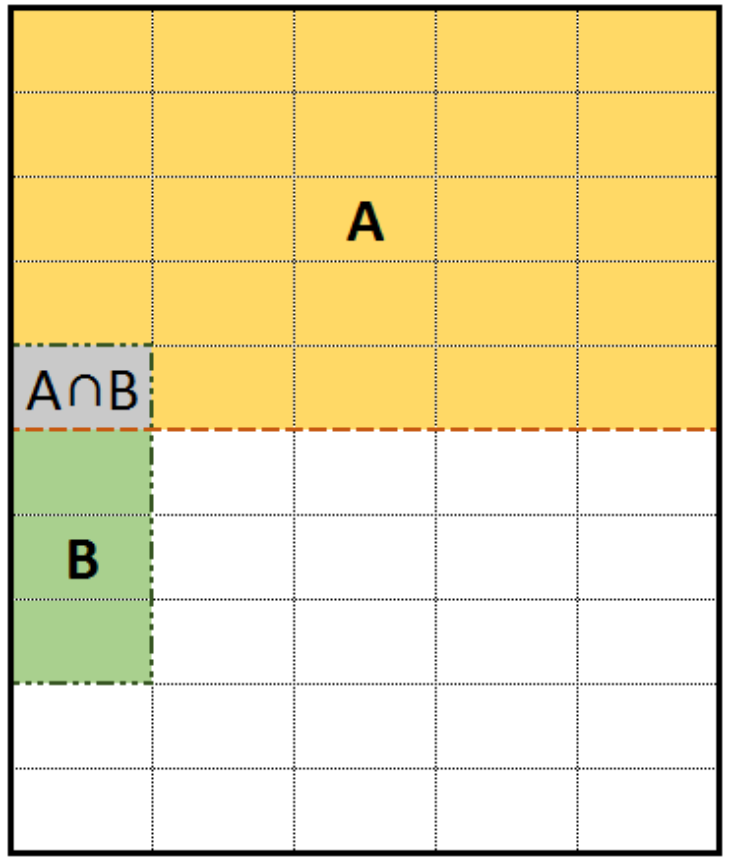
Observation	Clump Thickness	Uniformity of Cell Size	Bland Chromatin	Normal Nucleoli	Mitoses
1	3	1	3	1	1
2	6	8	3	7	1
3	4	1	3	1	1
4	2	1	3	1	1
5	1	1	3	1	1
86	1	2	1	1	1
87	3	1	1	1	1
88	4	2	2	1	1
89	1	1	2	1	1
90	4	3	3	3	1
136	1	1	2	1	1
137	1	1	1	1	1
138	3	1	2	1	2
139	2	1	1	1	1
140	4	8	10	6	1

Illustrating NBC With an Example

Theory – Conditional Probability

$P(A) = 0.5, P(B) = 0.08, P(A \cap B) = 0.02$

Probability that a randomly selected person belongs to both A and B is 2%.



Conditional probability of *A* given *B* is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.02}{0.08} = 0.25$$

Similarly,

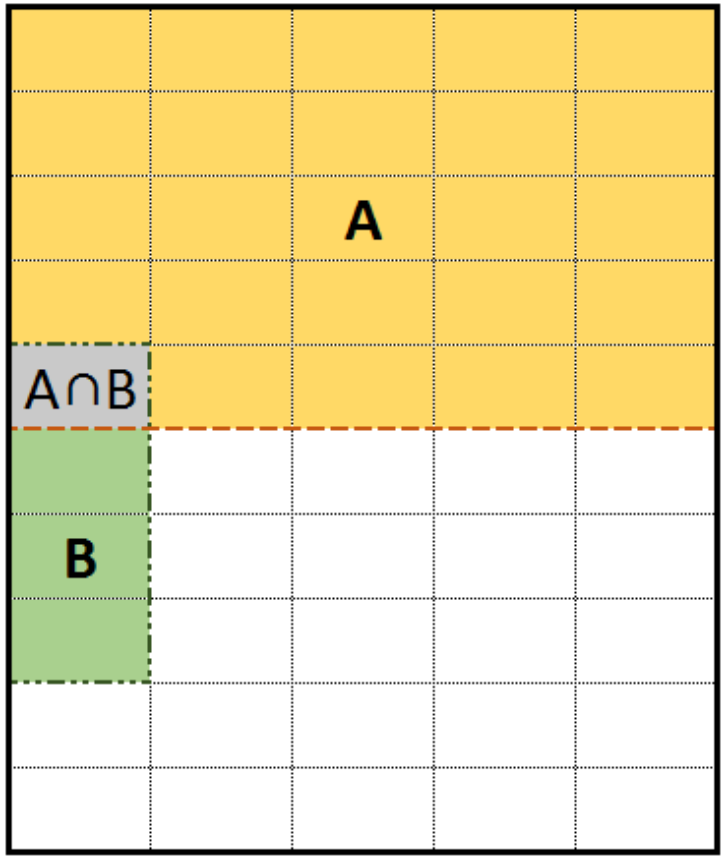
$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.02}{0.5} = 0.04$$

Illustrating NBC With an Example

Theory – Conditional Probability

Probability that a randomly selected person belongs to group *A* **knowing that they belong to group *B*** is 25%

Probability that a randomly selected person belongs to group *B* **knowing that they belong to group *A*** is 4%



Illustrating NBC With an Example

Theory – Naïve Bayes Classification

For each subject, the **Naïve Bayes Classifier** (NBC) compares the conditional probabilities

$$P(C_B|x_{i1}, x_{i2}, x_{i3}, \dots, x_{i9}) \text{ vs. } P(C_M|x_{i1}, x_{i2}, x_{i3}, \dots, x_{i9}),$$

where $x_{i1}, \dots, x_{i9}$ represent the following measurements:

Clump thickness	Bare nuclei
Uniformity of cell size	Bland chromatin
Uniformity of cell shape	Normal nucleoli
Marginal adhesion	Mitoses
Single epithelial cell size	

and classifies the tumor as benign or malignant based on the **largest probability**.

Illustrating NBC With an Example

Theory – Bayes' Theorem

The foundation of NBC rest on **Bayes' Theorem**, which can be expressed as

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{hypothesis}) \times P(\text{data}|\text{hypothesis})}{P(\text{data})}$$

This is often stated as

Posterior $\propto$ Prior $\times$ Likelihood

Illustrating NBC With an Example

Theory – Bayes' Theorem

In the context of the Breast Cancer dataset

- **posterior:** based on the collected data, how likely is it that a tumor is benign or malignant?
- **prior:** what are the proportions of benign and malignant tumors in general?
- **likelihood:** knowing that a tumor is benign (or malignant), how likely is it that its measurements would have been observed?

With our notation:

$$P(C_k|x_{i1}, \dots, x_{i9}) \propto P(C_k) \times P(x_{i1}, \dots, x_{i9}|C_k)$$

Illustrating NBC With an Example

Theory – NBC Assumptions

NBC requires only one assumption (which is why it is **naïve**):

For the individuals of a given class, the measurements are independent.

In our case, this means that for benign (or malignant) tumors, clump thickness is independent of uniformity of cell size, mitoses, and so on.

(Is this reasonable?)

Illustrating NBC With an Example

Theory – NBC Assumptions

With this assumption, the computation of the likelihood is **very** simple:

$$\text{Likelihood} = P(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{i9} | C_k)$$

$$= P(x_{i1} | C_k) P(x_{i2} | C_k) \cdots P(x_{i9} | C_k) = \prod_{j=1}^{9} P(x_{ij} | C_k)$$

The individual conditional probabilities are easier to compute.

Illustrating NBC With an Example

Theory – Steps required to perform NBC

Step 0: Preliminary analyses (visualization etc.)

Step 1: Calculate the prior probabilities

Step 2: Choose conditional distributions (for likelihoods)

Step 3: Estimate parameters (for likelihoods)

Step 4: Compute likelihoods

Step 5: Compute posterior probabilities

Step 6: Make classification decisions

Illustrating NBC With an Example

Step 1 – Calculate Prior Probabilities

What is the proportion of benign tumor?

- **Case 1.** We don't know much about them, so assume their proportions are equal

$$P(C_B) = P(C_M) = 50\%$$

- **Case 2*.** We can also ask the subject matter experts (SMEs), or base it on our training data:

$$P(C_B) = {}^{376}/_{559} = 67\%$$

$$P(C_M) = {}^{183}/_{559} = 33\%$$

Illustrating NBC With an Example

Step 2 – Choose Conditional Distributions

Case 1: Variables follow normal distributions

$$X_{ij}|C_k \sim N\left(\mu_k, \sigma_k^2\right)$$

Case 2*: Variables follow multinomial distribution

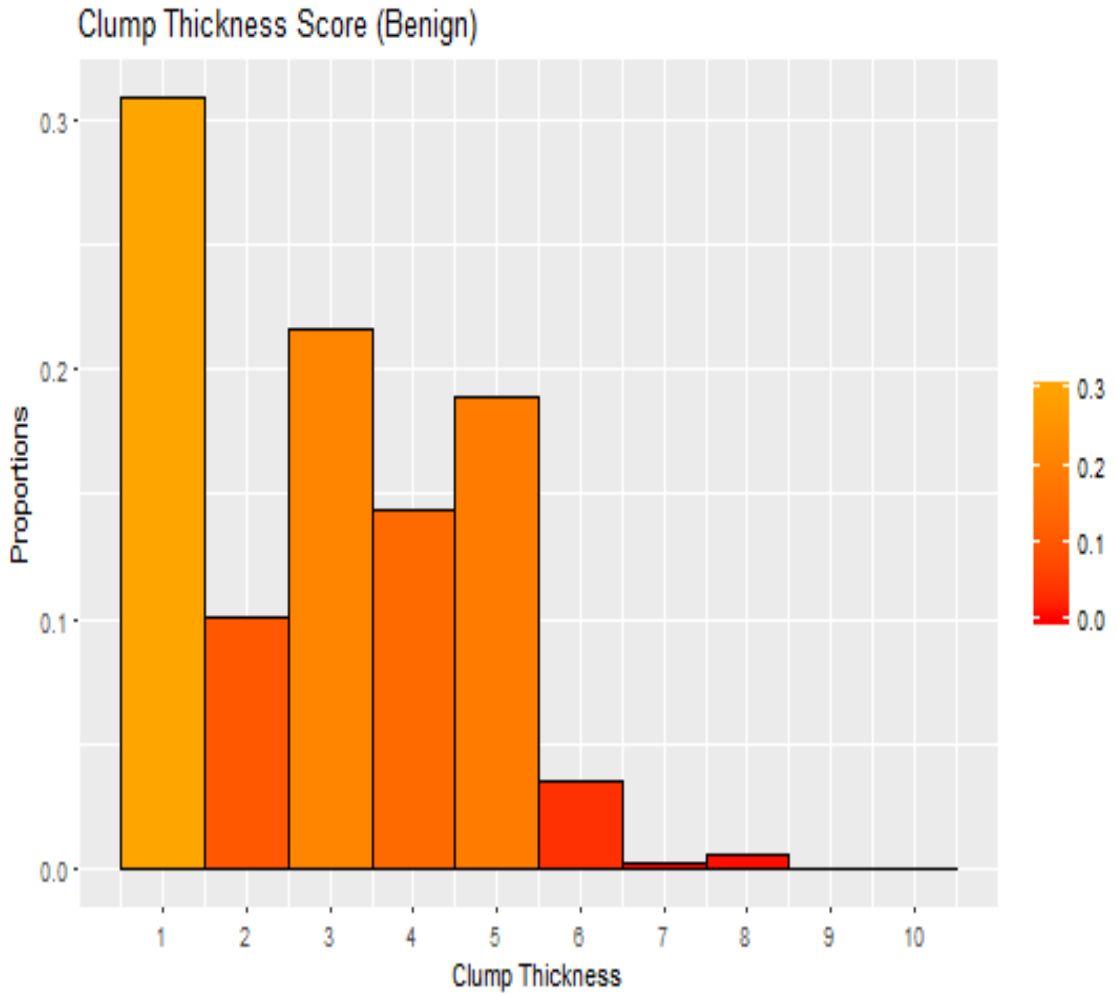
$$X_{ij}|C_k \sim \text{Multinomial}\left(p_{1,k}, p_{2,k}, \ldots, p_{d,k}\right)$$

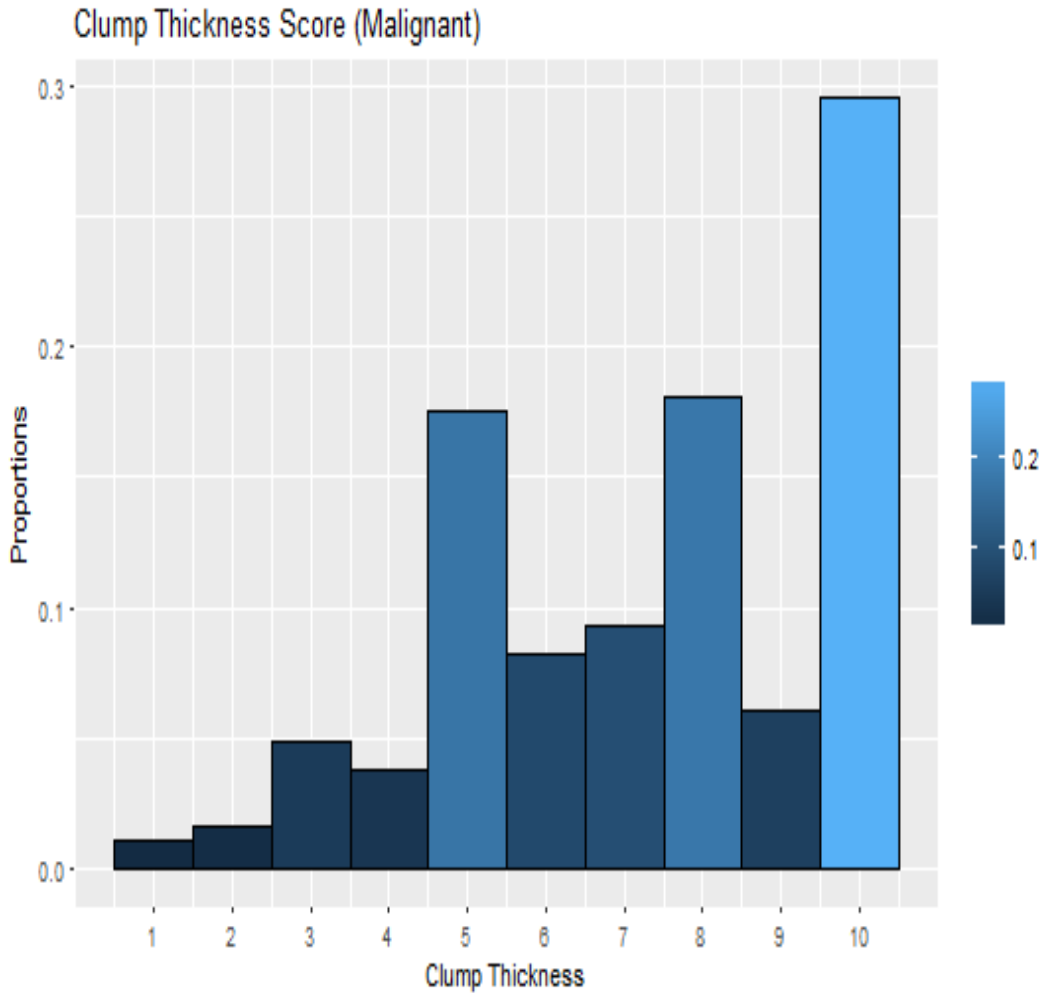
Given that the variables take values between 1 to 10, we use the multinomial distribution.

NOTE: These are not the only possibilities.

Illustrating NBC With an Example

Step 2 – Choose Conditional Distributions





Illustrating NBC With an Example

Step 3 – Estimate Parameters

In order to compute the likelihood, we need to estimate the parameters for the conditional distribution.

For each variable (e.g., clump thickness), we need 10 parameters: $p_1, \ldots, p_{10}$ (well, only 9…)

We estimate these *via* **observed proportions** in training data.

	Benign									Malignant								
Score	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	30.9%	83.2%	78.5%	81.9%	10.9%	85.4%	33.2%	88.0%	97.6%	1.1%	1.6%	1.1%	12.6%	0.5%	6.6%	1.1%	14.8%	50.8%
2	10.1%	7.7%	10.9%	8.5%	78.5%	4.3%	35.4%	6.9%	1.3%	1.6%	2.7%	3.3%	8.7%	9.8%	4.4%	3.3%	3.3%	13.1%
3	21.5%	6.6%	6.1%	6.4%	6.9%	2.9%	27.4%	2.4%	0.3%	4.9%	11.5%	8.7%	12.0%	16.4%	6.0%	16.4%	10.4%	14.8%
4	14.4%	1.6%	2.9%	1.1%	1.6%	1.3%	1.6%	0.3%	0.0%	3.8%	13.1%	13.7%	10.4%	14.2%	4.4%	12.6%	7.1%	6.0%
5	18.9%	0.0%	0.3%	0.8%	0.8%	2.4%	0.8%	0.5%	0.3%	17.5%	10.9%	12.0%	8.7%	16.9%	8.2%	14.2%	8.7%	2.2%
6	3.5%	0.3%	0.8%	0.8%	0.5%	0.0%	0.3%	0.8%	0.0%	8.2%	10.9%	9.8%	6.6%	15.3%	2.2%	2.7%	7.7%	1.1%
7	0.3%	0.3%	0.5%	0.0%	0.3%	0.0%	1.3%	0.3%	0.3%	9.3%	8.2%	12.0%	5.5%	4.9%	3.3%	25.7%	4.9%	3.3%
8	0.5%	0.0%	0.0%	0.0%	0.5%	0.5%	0.0%	0.8%	0.3%	18.0%	12.0%	12.6%	10.4%	8.7%	7.1%	12.6%	9.8%	3.3%
9	0.0%	0.3%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	6.0%	2.7%	2.2%	1.1%	0.5%	4.4%	4.9%	6.6%	0.0%
10	0.0%	0.0%	0.0%	0.3%	0.0%	0.8%	0.0%	0.0%	0.0%	29.5%	26.2%	24.6%	24.0%	12.6%	53.0%	6.6%	26.8%	5.5%

Illustrating NBC With an Example

Step 4 – Compute the Likelihood

The first test subject has following observation

Obs.	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	3	1	1	1	2	2	3	1	1

For each of the **Benign** and **Malignant** cases, the probabilities for each test score are highlighted. Naïve assumption says the likelihood is obtained by **multiplying the probabilities.**

	Benign									Malignant								
Score	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	30.9%	83.2%	78.5%	81.9%	10.9%	85.4%	33.2%	88.0%	97.6%	1.1%	1.6%	1.1%	12.6%	0.5%	6.6%	1.1%	14.8%	50.8%
2	10.1%	7.7%	10.9%	8.5%	78.5%	4.3%	35.4%	6.8%	1.3%	1.6%	2.7%	3.3%	8.7%	9.8%	4.4%	3.3%	3.3%	13.1%
3	21.5%	6.6%	6.1%	6.4%	6.9%	2.9%	27.4%	2.4%	0.3%	4.9%	11.5%	8.7%	12.0%	16.4%	6.0%	16.4%	10.4%	14.3%
4	14.4%	1.6%	2.9%	1.1%	1.6%	1.3%	1.6%	0.3%	0.0%	3.8%	13.1%	13.7%	10.4%	14.2%	4.4%	12.6%	7.1%	6.0%
5	18.9%	0.0%	0.3%	0.8%	0.8%	2.4%	0.8%	0.5%	0.3%	17.5%	10.9%	12.0%	8.7%	16.9%	8.2%	14.2%	8.7%	2.2%
6	3.5%	0.3%	0.8%	0.8%	0.5%	0.0%	0.3%	0.8%	0.0%	8.2%	10.9%	9.8%	6.6%	15.3%	2.2%	3.7%	7.7%	1.1%
7	0.3%	0.3%	0.5%	0.0%	0.3%	0.0%	1.3%	0.3%	0.3%	9.3%	8.2%	12.0%	5.5%	4.3%	3.3%	25.7%	4.9%	3.3%
8	0.5%	0.0%	0.0%	0.0%	0.5%	0.5%	0.0%	0.8%	0.3%	18.0%	12.0%	12.6%	10.4%	8.7%	7.1%	12.6%	9.8%	3.3%
9	0.0%	0.3%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.7%	2.2%	1.1%	0.5%	4.4%	4.3%	6.6%	0.0%
10	0.0%	0.0%	0.0%	0.3%	0.0%	0.8%	0.0%	0.0%	0.0%	29.5%	26.2%	24.6%	24.0%	12.6%	53.0%	6.8%	26.8%	5.5%
Likelihood										9.06E-04								5.85E-11

Illustrating NBC With an Example

Step 5 – Compute Posterior Probabilities

Recall that

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

Class	Prior	Likelihood	Posterior
Benign	6.73E-01	9.06E-04	6.09E-04
Malignant	3.27E-01	5.85E-11	1.92E-11

Based on NBC, we believe that the tumor on the first test subject is **benign** (and classify it as such).

Illustrating NBC With an Example

Step 6 – Conclude Based on a Decision Rule

Similarly, we can calculate the posterior probabilities for each case

Based on NBC, we classify

- $^{78}/_{140} = 55.7\%$ as **benign**
- $^{62}/_{140} = 44.3\%$ as **malignant**

How well did NBC perform for the classification?

Observation	Posterior (numerator)	
	Benign	Malignant
1	6.09E-04	1.92E-11
2	0.00E+00	1.50E-09
3	6.35E-04	2.14E-11
4	7.97E-04	2.87E-11
5	6.00E-04	5.85E-12
86	1.53E-04	5.68E-12
87	1.48E-02	1.92E-12
88	1.35E-04	2.24E-11
89	2.26E-02	1.28E-12
90	1.38E-06	8.81E-10
136	3.14E-03	3.83E-12
137	1.66E-03	4.07E-13
138	2.15E-04	1.48E-12
139	6.96E-03	6.39E-13
140	0.00E+00	2.81E-10

Illustrating NBC With an Example
Debriefing – Performance Evaluation

		Prediction		
		Benign	Malignant	
Truth	Benign	77	5	82
	Malignant	1	57	58
		78	62	140

Overall misclassification rate: $^{6}/_{140} = 4.3\%$

(False positive) Benign misclassification rate: $^{5}/_{82} = 6.1\%$

(False negative) Malignant misclassification rate: $^{1}/_{58} = 1.7\%$

Malignant detection is more accurate than benign detection.

Illustrating NBC With an Example
Debriefing – Performance Evaluation

ID	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	True class	Posterior	
											Benign	Malignant
2	6	8	8	1	3	4	3	7	1	Benign	0	1.50E-09
23	1	1	1	1	10	1	1	1	1	Benign	0	5.44E-13
33	8	4	4	5	4	7	7	8	2	Benign	0	1.43E-09
61	4	6	5	6	7	0	4	9	1	Benign	0	2.22E-09
74	3	4	5	3	7	3	4	6	1	Benign	3.80E-15	4.41E-10
102	6	3	2	1	3	4	4	1	1	Malignant	1.74E-09	8.57E-10

First, notice that some of the posterior probabilities are 0

- This is because we did not observe all their scores in the training set

Also, take a look at ID 23. Is anything out of the ordinary there?

NOTE: a score of 0 means that the observation was missing

Illustrating NBC With an Example

Debriefing – Performance Evaluation

ID	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	True class	Posterior	
											Benign	Malignant
2	6	8	8	1	3	4	3	7	1	Benign	0	1.50E-09
23	1	1	1	1	10	1	1	1	1	Benign	0	5.44E-13
33	8	4	4	5	4	7	7	8	2	Benign	0	1.43E-09
61	4	6	5	6	7	0	4	9	1	Benign	0	2.22E-09
74	3	4	5	3	7	3	4	6	1	Benign	3.80E-15	4.41E-10
102	6	3	2	1	3	4	4	1	1	Malignant	1.74E-09	8.57E-10

	Benign								
Score	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
1	30.9%	83.2%	78.5%	81.9%	10.9%	85.4%	33.2%	88.0%	97.6%
2	10.1%	7.7%	10.9%	8.5%	78.5%	4.3%	35.4%	6.9%	1.3%
3	21.5%	6.6%	6.1%	6.4%	6.9%	2.9%	27.4%	2.4%	0.3%
4	14.4%	1.6%	2.9%	1.1%	1.6%	1.3%	1.6%	0.3%	0.0%
5	18.9%	0.0%	0.3%	0.8%	0.8%	2.4%	0.8%	0.5%	0.3%
6	3.5%	0.3%	0.8%	0.8%	0.5%	0.0%	0.3%	0.8%	0.0%
7	0.3%	0.3%	0.5%	0.0%	0.3%	0.0%	1.3%	0.3%	0.3%
8	0.5%	0.0%	0.0%	0.0%	0.5%	0.5%	0.0%	0.8%	0.3%
9	0.0%	0.3%	0.0%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%
10	0.0%	0.0%	0.0%	0.3%	0.0%	0.8%	0.0%	0.0%	0.0%

Illustrating NBC With an Example
Debriefing – Performance Evaluation

In general, we have

- **Benign** tumors = lower scores
- **Malignant** tumors = higher scores

And if we do not observe a pattern in the training set, the resulting posterior probability will always be 0.

In case of ID 23, only the score from the single epithelial cell size is abnormal, and all other scores are indicating it to be benign.

How can we avoid this issue of posterior probability = 0?

Illustrating NBC With an Example

Debriefing – Alternative NBC

Set a base probability, say 10^{-8}

Multiply it by the smallest non-zero probability (0.3%), which gives $3\times10^{-11}\%$.

Replace the zero probability by that result: $3\times10^{-11}\%$.

Now, the probabilities add up to $1+3\times10^{-13}$, so we normalize them before proceeding to steps 4 – 6.

Original		Step 1		Step 2
Single Epithelial Cell Size		Single Epithelial Cell Size		Single Epithelial Cell Size
10.9%		10.9%		10.9%
78.5%		78.5%		78.5%
6.9%		6.9%		6.9%
1.6%	⇒	1.6%	⇒	1.6%
0.8%		0.8%		0.8%
0.5%		0.5%		0.5%
0.3%		0.3%		0.3%
0.5%		0.5%		0.5%
0.0%		3E-11%		3E-11%
0.0%		3E-11%		3E-11%

Illustrating NBC With an Example
Debriefing – Alternative NBC

By assigning non-zero prob. for all scores, the total error rate decreased from $^{6}/_{140}$ to $^{5}/_{140}$.

		Prediction (original)		
		Benign	Malignant	
Truth	Benign	77	5	82
	Malignant	1	57	58
		78	62	140

⇒

		Prediction (alternative)		
		Benign	Malignant	
Truth	Benign	78	4	82
	Malignant	1	57	58
		79	61	140

Posterior probabilities are all non-zero; however, most of them are very small.

ID 23 is now correctly classified as **Benign** even with strong penalty for unusual observation (possible recording error?).

ID	True class	Posterior (original)			Posterior (alternative)	
		Benign	Malignant		Benign	Malignant
2	Benign	0	1.50E-09		9.03E-30	1.50E-09
23	Benign	0	5.44E-13	⇒	7.38E-13	5.47E-13
33	Benign	0	1.43E-09		1.64E-26	1.43E-09
61	Benign	0	2.22E-09		6.01E-24	2.22E-09
74	Benign	3.80E-15	4.41E-10		3.89E-15	4.43E-10
102	Malignant	1.74E-09	8.57E-10		1.78E-09	8.62E-10

Notes and Comments

Independence Assumption and Robustness

The variables are **NOT** independent; but NBC still works well with test data ⇒ Method may be robust against departure from the independence assumption.

Dependency among variable may change the posterior probabilities, but the class with maximum posterior probability is often unchanged (Domingos and Pazzani, 1997).

In the classification context, we typically get more insight from independent (or uncorrelated) data than from correlated data.