The Curse of Dimensionality

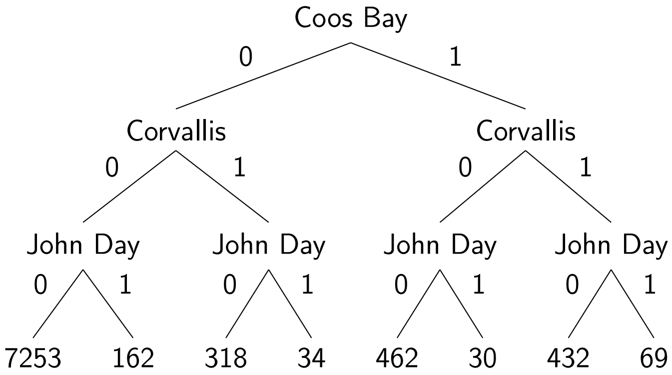
and Naive Bayes

Probability and Statistics for Data Science

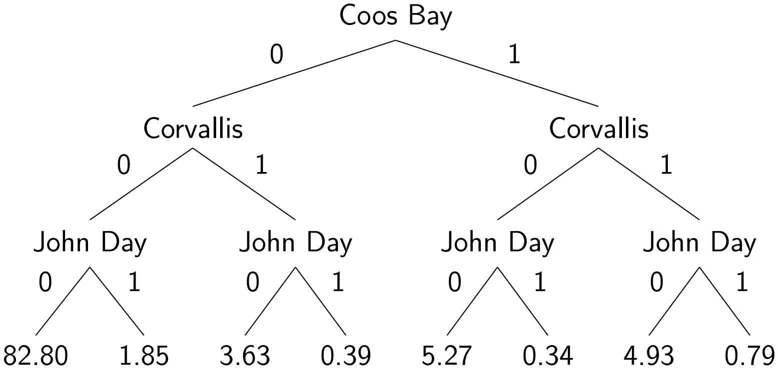
Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

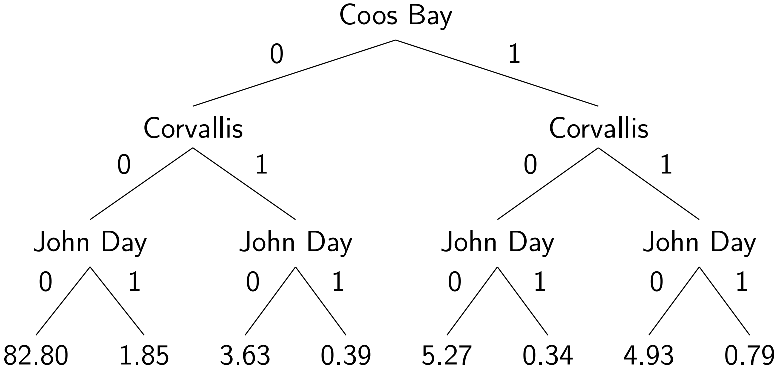
Precipitation in Oregon



Empirical joint pmf (%)



What if we have 134 stations?



US weather dataset

Number of weather stations: 134

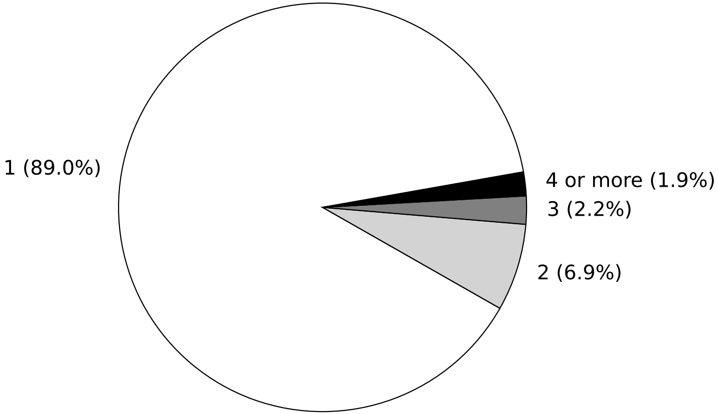
Number of possible patterns: $2^{134} \geq 10^{40}$!!!

Number of data: 8,760...

Dependencies explode exponentially: This is the curse of dimensionality!

Maybe a few patterns are repeated very often?

No...



Possible solution

Assume independence

If all stations are independent, number of parameters? 134

But we are not modeling dependencies...

Pragmatic compromise: Conditional dependence assumptions

Classification

Data: (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n)

x_i is d-dimensional vector (e.g. picture), y_i is class (e.g. *dog*)

Goal: Assign class to new data

Probabilistic modeling

Model data as random vector $\tilde{x}$ and class as random variable $\tilde{y}$

For new data vector x:

$$\hat{y} := \arg \max_{y \in \{1,2,\ldots,c\}} p_{\tilde{y} \,|\, \tilde{x}}(y \,|\, x)$$

Predicting political affiliation

U.S. House of Representatives in 1984

Goal: Predict whether politician is Republican or Democrat based on voting record

Training set: 425 politicians

Test set: 10 politicians

Probabilistic model

We model affiliation with the random variable

$$\tilde{y} = \begin{cases} R & \text{if representative is a Republican} \\ D & \text{if representative is a Democrat} \end{cases}$$

and votes as a 16-dimensional random vector $\tilde{x}$

$$\tilde{x}[i] = \begin{cases} 1 & \text{if representative voted Yes on issue } i \\ 0 & \text{otherwise} \end{cases}$$

Goal: Estimate $p_{\tilde{y} \mid \tilde{x}}(\cdot \mid x)$ for any x

Possible values of x? $2^{16} = 65{,}536$

Training data points: 425...

Naive assumption

Can we assume everything is independent? Not $\tilde{x}$ and $\tilde{y}$!

We assume votes are conditionally independent given affiliation

$$p_{\tilde{x} \mid \tilde{y}}(x \mid R) = \prod_{i=1}^{d} p_{\tilde{x}[i] \mid \tilde{y}}(x[i] \mid R)$$

$$p_{\tilde{x} \mid \tilde{y}}(x \mid D) = \prod_{i=1}^{d} p_{\tilde{x}[i] \mid \tilde{y}}(x[i] \mid D)$$

What we really want is $p_{\tilde{y} \mid \tilde{x}}$

Bayes rule

$$
\begin{aligned}
&p_{\tilde{y}\,|\,\tilde{x}}(R\,|\,x) \\
&= \frac{p_{\tilde{y},\tilde{x}}(R,x)}{p_{\tilde{x}}(x)} \\
&= \frac{p_{\tilde{y}}(R)p_{\tilde{x}\,|\,\tilde{y}}(x\,|\,R)}{p_{\tilde{y},\tilde{x}}(R,x)+p_{\tilde{y},\tilde{x}}(D,x)} \\
&= \frac{p_{\tilde{y}}(R)\prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,R)}{p_{\tilde{y}}(R)\prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,R)+p_{\tilde{y}}(D)\prod_{i=1}^{d} p_{\tilde{x}[i]\,|\,\tilde{y}}(x[i]\,|\,D)}
\end{aligned}
$$

How many parameters do we need to estimate?

Model

$p_{\tilde{y}}(R) = 0.381 \qquad (p_{\tilde{y}}(D) = 0.619)$

i	1	2	3	4	5	6	7	8
$p_{\tilde{x}[i]\,\vert\,\tilde{y}}(1\,\vert\,R)$	0.19	0.50	0.14	0.99	0.95	0.90	0.24	0.15
$p_{\tilde{x}[i]\,\vert\,\tilde{y}}(1\,\vert\,D)$	0.61	0.50	0.89	0.05	0.22	0.47	0.78	0.83

i	9	10	11	12	13	14	15	16
$p_{\tilde{x}[i]\,\vert\,\tilde{y}}(1\,\vert\,R)$	0.11	0.55	0.14	0.87	0.86	0.98	0.09	0.66
$p_{\tilde{x}[i]\,\vert\,\tilde{y}}(1\,\vert\,D)$	0.76	0.47	0.51	0.15	0.29	0.35	0.64	0.94

Parameters: $16 \cdot 2 + 1 = 33$

Applying the model

i	1	2	3	4	5	6	7	8
$p_{\tilde{x}[i] \mid \tilde{y}}(1 \mid R)$	0.19	0.50	0.14	0.99	0.95	0.90	0.24	0.15
$p_{\tilde{x}[i] \mid \tilde{y}}(1 \mid D)$	0.61	0.50	0.89	0.05	0.22	0.47	0.78	0.83
Example	N	–	**Y**	**N**	**N**	Y	**Y**	**Y**

i	9	10	11	12	13	14	15	16
$p_{\tilde{x}[i] \mid \tilde{y}}(1 \mid R)$	0.11	0.55	0.14	0.87	0.86	0.98	0.09	0.66
$p_{\tilde{x}[i] \mid \tilde{y}}(1 \mid D)$	0.76	0.47	0.51	0.15	0.29	0.35	0.64	0.94
Example	N	Y	N	**N**	**N**	**N**	**Y**	–

$$
\begin{aligned}
&p_{\tilde{y} \mid \tilde{x}}(D \mid x) \\
&= \frac{p_{\tilde{y}}(D) \prod_{i \in \{1,3,\ldots,15\}} p_{\tilde{x}[i] \mid \tilde{y}}(x[i] \mid D)}{p_{\tilde{y}}(D) \prod_{i \in \{1,3,\ldots,15\}} p_{\tilde{x}[i] \mid \tilde{y}}(x[i] \mid D) + p_{\tilde{y}}(R) \prod_{i \in \{1,3,\ldots,15\}} p_{\tilde{x}[i] \mid \tilde{y}}(x[i] \mid R)} \\
&= 1 - 1.4 10^{-8}
\end{aligned}
$$

9/10 correct predictions on test data

What have we learned?

Estimating joint pmfs is often impossible due to curse of dimensionality

Conditional independence assumptions can help

Classification via naive Bayes