

Multiple Discrete Variables (Overview)

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>

Key questions

How to jointly model multiple uncertain quantities

How to estimate causal relationships from data

How to fight the curse of dimensionality

Question 1

How to jointly model multiple uncertain quantities

Represent them as random variables in the **same probability space**

Rolling a die twice

Probability space representing two rolls of a six-sided die

Outcomes:

$$\omega := \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} \quad \omega_1, \omega_2 \in \{1, 2, 3, 4, 5, 6\}$$

Random variables

Define random variables to represent first roll, second roll and sum of rolls

$$\tilde{a}(\omega) := \omega_1$$

$$\tilde{b}(\omega) := \omega_2$$

$$\tilde{c}(\omega) := \omega_1 + \omega_2$$

The outcome **fixes** the values of all random variables **simultaneously**

$$\text{If } \omega = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \quad \tilde{a}(\omega) = 3 \quad \tilde{b}(\omega) = 1 \quad \tilde{c}(\omega) = 4$$

Sample space

| | |
|------------------------------|------------------------------|
| $A_1 := \{\tilde{a} = a_1\}$ | $A_2 := \{\tilde{a} = a_2\}$ |
|------------------------------|------------------------------|

Ω

| |
|------------------------------|
| $B_1 := \{\tilde{b} = b_1\}$ |
| $B_2 := \{\tilde{b} = b_2\}$ |

Ω

| | |
|----------------|----------------|
| $A_1 \cap B_1$ | $A_2 \cap B_1$ |
| $A_1 \cap B_2$ | $A_2 \cap B_2$ |

Ω

Sample space

| | |
|------------------------------|------------------------------|
| $A_1 := \{\tilde{a} = a_1\}$ | $A_2 := \{\tilde{a} = a_2\}$ |
|------------------------------|------------------------------|

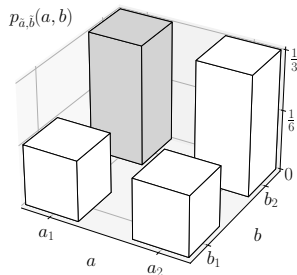
Ω

| |
|------------------------------|
| $B_1 := \{\tilde{b} = b_1\}$ |
| $B_2 := \{\tilde{b} = b_2\}$ |

Ω

| | |
|----------------|----------------|
| $A_1 \cap B_1$ | $A_2 \cap B_1$ |
| $A_1 \cap B_2$ | $A_2 \cap B_2$ |

Ω

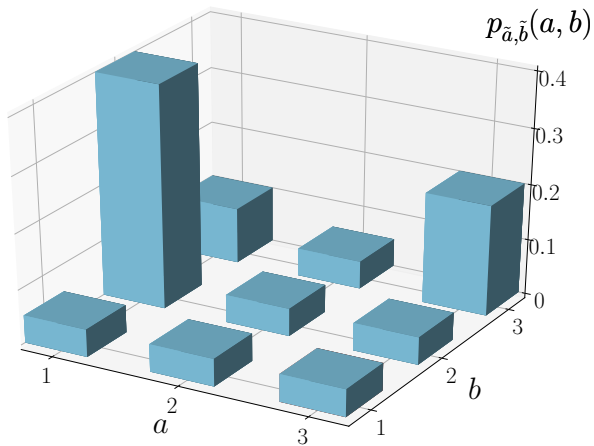


Joint probability mass function

The joint pmf of $\tilde{a} : \Omega \rightarrow A$ and $\tilde{b} : \Omega \rightarrow B$ is defined as

$$p_{\tilde{a}, \tilde{b}}(a, b) := \mathbb{P}(\tilde{a} = a, \tilde{b} = b)$$

Joint pmf



Random vector

Each entry $\tilde{x}[i]$ is a random variable in the same probability space

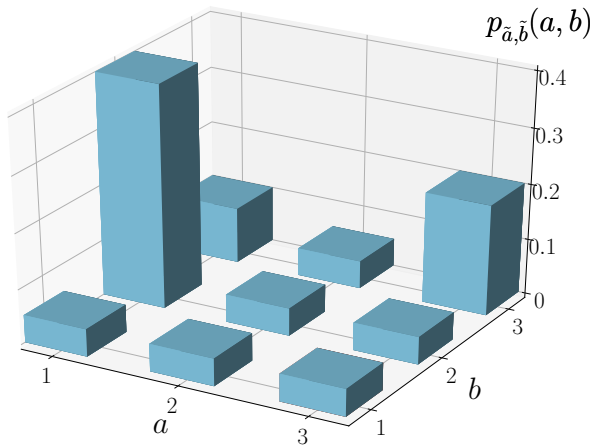
$$\tilde{x} := \begin{bmatrix} \tilde{x}[1] \\ \tilde{x}[2] \\ \dots \\ \tilde{x}[d] \end{bmatrix}$$

Joint probability mass function

The joint pmf of a discrete random vector \tilde{x} is

$$p_{\tilde{x}}(x) := P(\tilde{x}[1] = x[1], \tilde{x}[2] = x[2], \dots, \tilde{x}[d] = x[d])$$

Computing probabilities



$$P(\{\tilde{a} < 2, \tilde{b} > 1\}) = p_{\tilde{a}, \tilde{b}}(1, 2) + p_{\tilde{a}, \tilde{b}}(1, 3) = 0.5$$

Properties of joint pmfs?

Joint pmfs are nonnegative (they are probabilities)

$$\sum_{a \in A} \sum_{b \in B} p_{\tilde{a}, \tilde{b}}(a, b) = \mathbb{P} \left(\{\tilde{a} \in A\} \cap \{\tilde{b} \in B\} \right) = 1$$

$$\sum_{x[1] \in R_1} \sum_{x[2] \in R_2} \cdots \sum_{x[d] \in R_d} p_{\tilde{x}}(x) = 1$$

Any function with these properties is a **valid joint pmf**

Estimating a joint pmf from data

If data equal

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

How would you estimate $p_{\tilde{a}, \tilde{b}}(\begin{bmatrix} 2 \\ 1 \end{bmatrix})$?

Empirical joint pmf

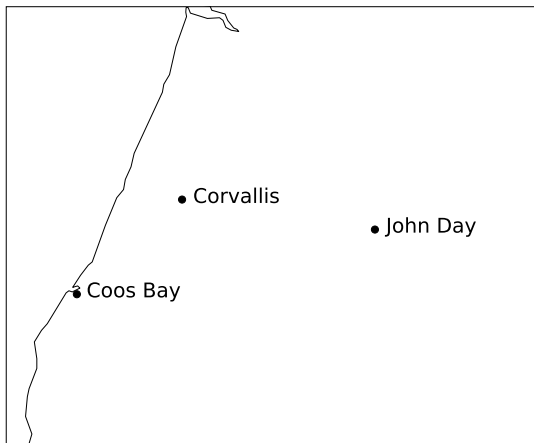
Data: $X := \{x_1, x_2, \dots, x_n\}$

The empirical joint pmf is

$$p_X(v) := \frac{\sum_{i=1}^n 1_{x_i=v}}{n},$$

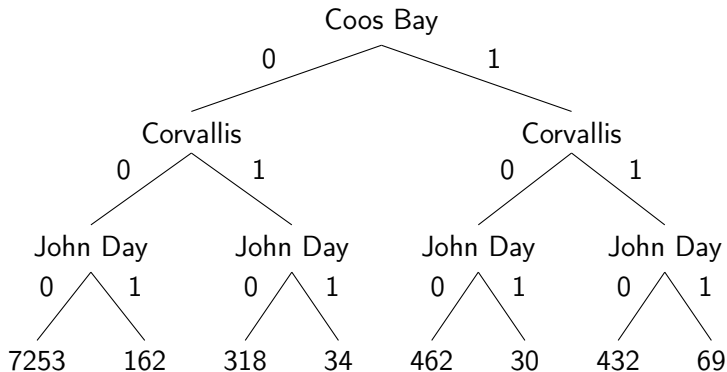
where $1_{x_i=v}$ equals one if $x_i = v$ and zero otherwise

Precipitation in Oregon

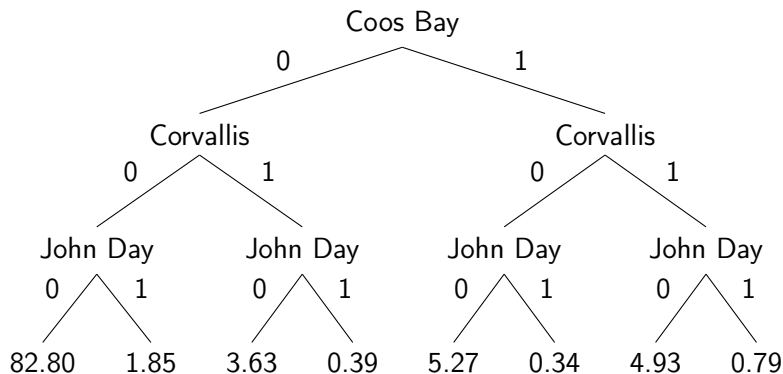


Goal: Model precipitation in Coos Bay, Corvallis, John Day

Precipitation in Oregon



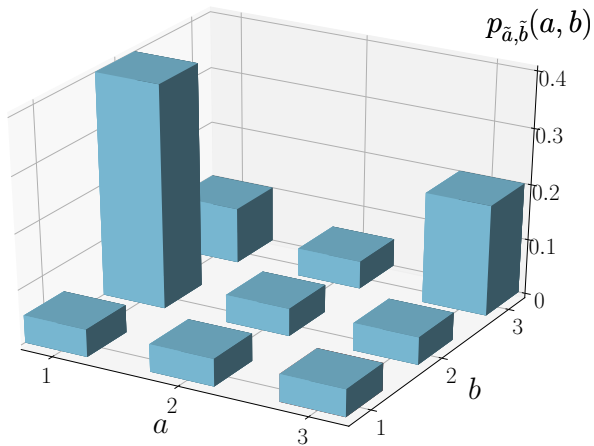
Empirical joint pmf (%)



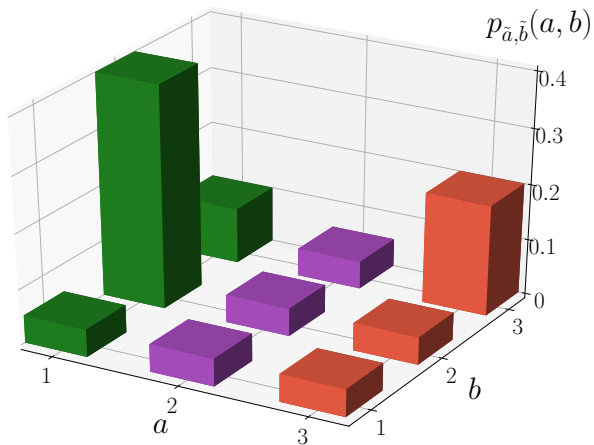
Marginal distributions

In a model with many variables, how do we characterize behavior of individual variables?

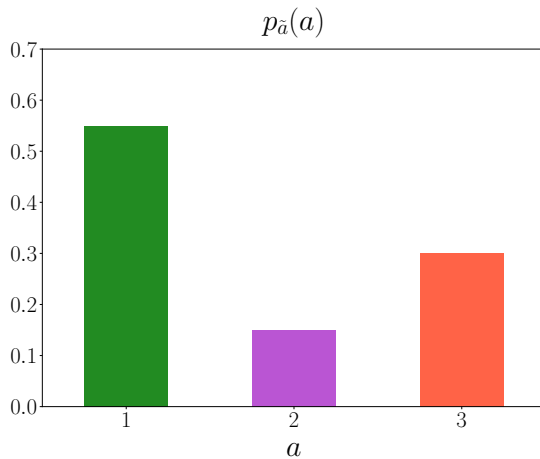
$p_{\tilde{a}}?$



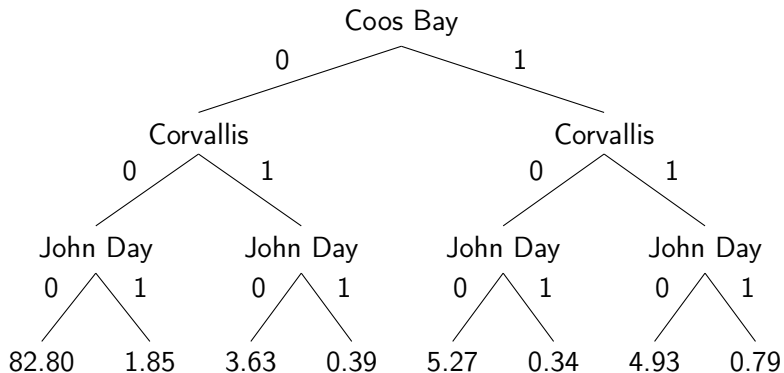
Marginal pmf



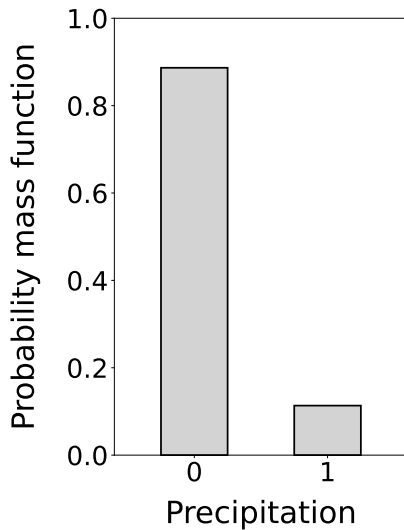
Marginal pmf



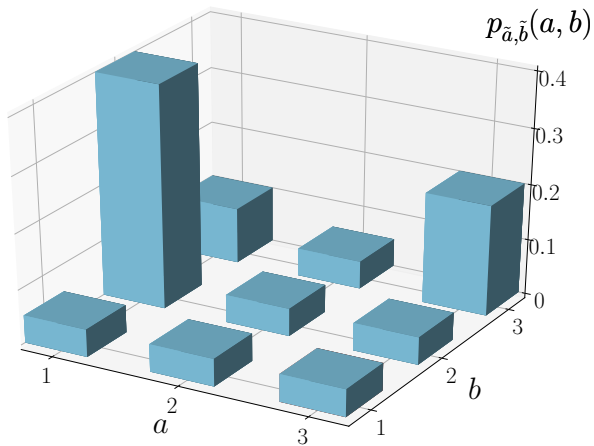
Coos Bay?



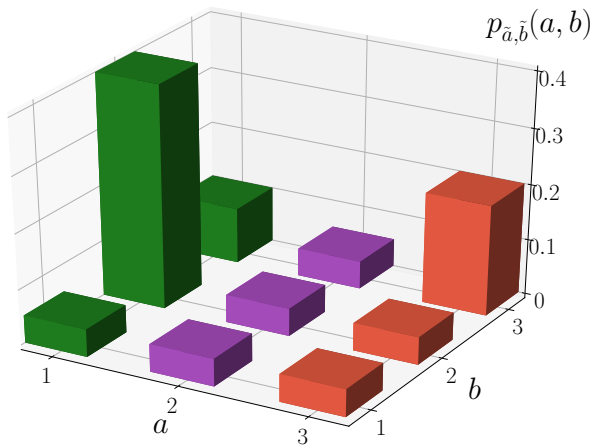
Marginal pmf



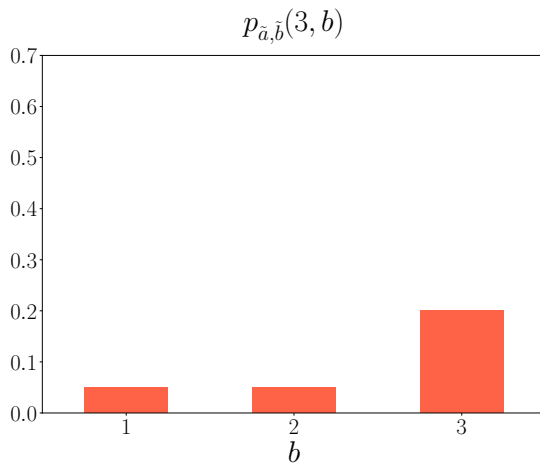
What if we know that $\tilde{a} = 3$?



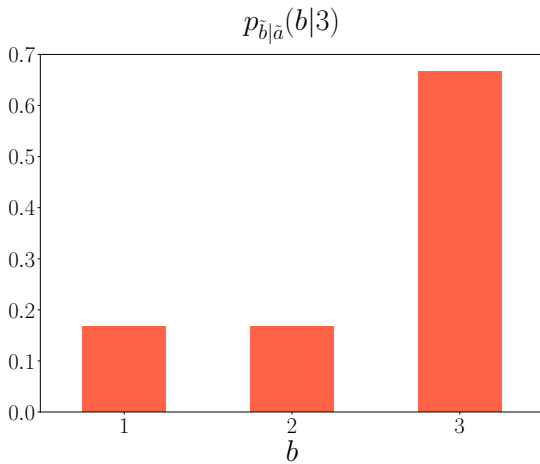
What if we know that $\tilde{a} = 3$?



Is this a valid pmf?



$$p_{\tilde{b}|\tilde{a}}(b|a) = \frac{p_{\tilde{a},\tilde{b}}(a,b)}{p_{\tilde{a}}(a)}$$



Chain rule for discrete random variables

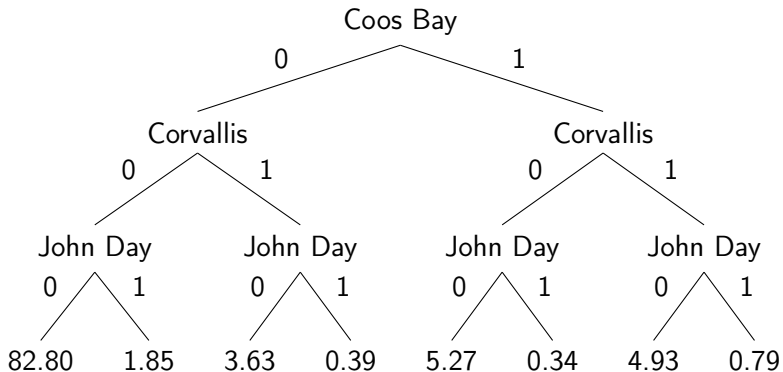
$$\begin{aligned} p_{\tilde{a}, \tilde{b}}(a, b) &= p_{\tilde{a}}(a) p_{\tilde{b} | \tilde{a}}(b | a) \\ &= p_{\tilde{b}}(b) p_{\tilde{a} | \tilde{b}}(a | b) \end{aligned}$$

Chain rule for discrete random vectors

$$p_{\tilde{x}}(x) = p_{\tilde{x}[1]}(x[1]) \prod_{i=1}^n p_{\tilde{x}[i] \mid \tilde{x}[1], \dots, \tilde{x}[i-1]}(x[i] \mid x[1], \dots, x[i-1])$$

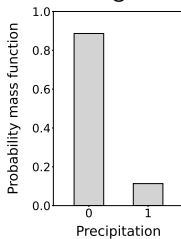
Any order works!

Corvallis = 0, John Day = 0?

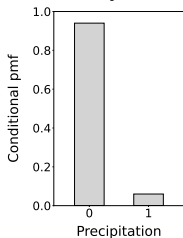


Conditional pmfs

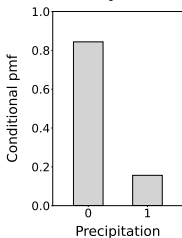
Marginal



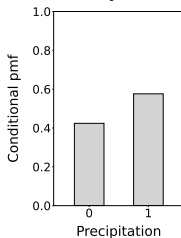
Corvallis = 0
John Day = 0



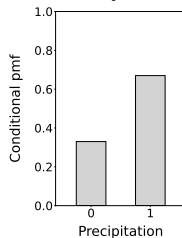
Corvallis = 0
John Day = 1



Corvallis = 1
John Day = 0



Corvallis = 1
John Day = 1

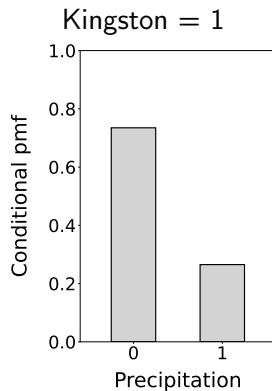
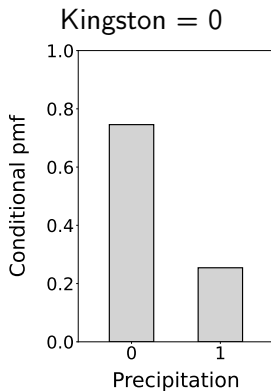
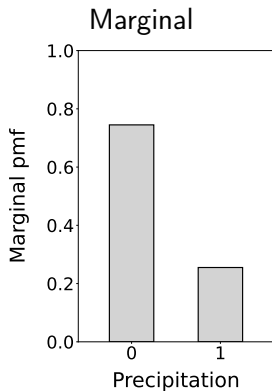


Precipitation in Kingston and Hilo



Goal: Model precipitation in Kingston and Hilo

Marginal and conditional pmfs



Intuition

Two random variables \tilde{a} and \tilde{b} are **independent** if our uncertainty about \tilde{a} does **not** change when information about \tilde{b} is revealed

Independence

\tilde{a} and \tilde{b} are independent if for any a and b

$$p_{\tilde{a}|\tilde{b}}(a|b) = P(\tilde{a} = a | \tilde{b} = b) = P(\tilde{a} = a) = p_{\tilde{a}}(a)$$

Equivalently,

$$p_{\tilde{a},\tilde{b}}(a,b) = p_{\tilde{a}}(a)p_{\tilde{b}}(b)$$

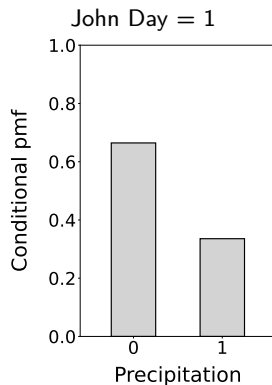
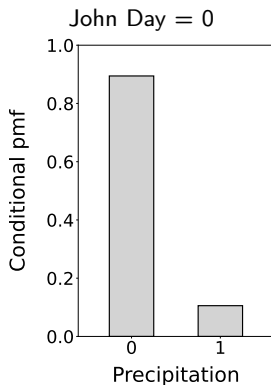
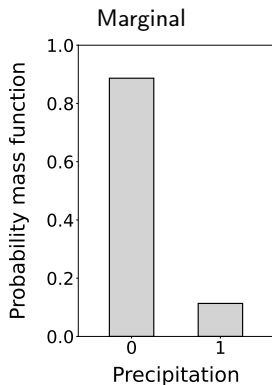
Independence

The d entries $\tilde{x}[1], \tilde{x}[2], \dots, \tilde{x}[d]$ in a discrete random vector \tilde{x} are mutually independent if and only if

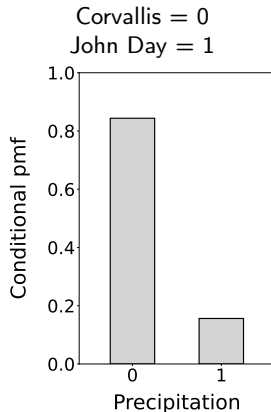
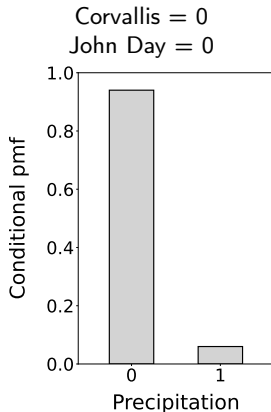
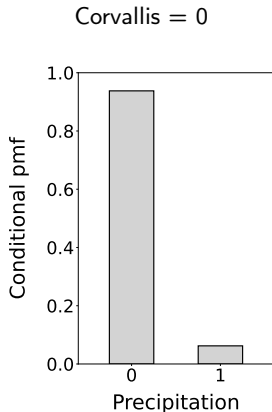
$$p_{\tilde{x}}(x) = \prod_{i=1}^d p_{\tilde{x}[i]}(x[i])$$

for all possible values of the entries

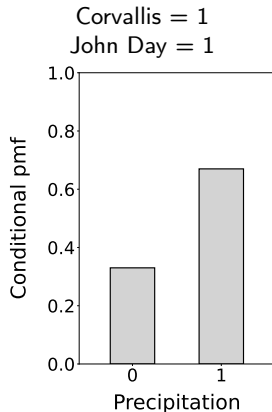
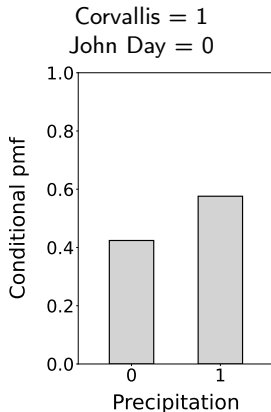
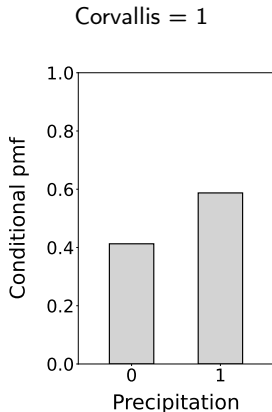
Coos Bay and John Day



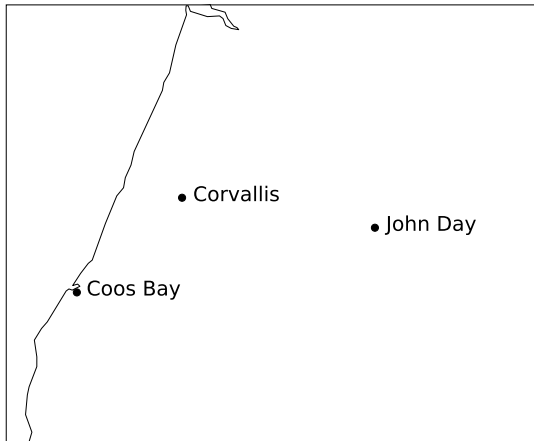
Coos Bay given Corvallis and John Day



Coos Bay given Corvallis and John Day



Precipitation in Oregon



Conditional independence

Two random variables \tilde{a} and \tilde{b} are **conditionally** independent given \tilde{c} if our uncertainty about \tilde{a} does **not** change when \tilde{b} is revealed, **as long as the value of \tilde{c} is known**

Conditional independence

\tilde{a} and \tilde{b} are conditionally independent given \tilde{c} if

$$p_{\tilde{a}, \tilde{b} | \tilde{c}}(a, b | c) = p_{\tilde{a} | \tilde{c}}(a | c) p_{\tilde{b} | \tilde{c}}(b | c) \quad \text{for all } a, b, c$$

Question 2

How to estimate causal relationships from data

Using conditional probabilities, but being **very careful**

Data

NBA games from the 2014/2015 season

3-point shot percentage

Stephen Curry: 41.7%

Courtney Lee: **43.9%**

Was Lee the better shooter?

A closer look at the data

| | Stephen Curry | Courtney Lee |
|--------------------------------|-----------------------------|----------------------------|
| Short threes (≤ 24 feet) | $45/90 = \mathbf{50.0\%}$ | $56/116 = 48.3\%$ |
| Long threes (> 24 feet) | $145/366 = \mathbf{39.6\%}$ | $19/55 = 34.5\%$ |
| Total | $190/456 = 41.7\%$ | $75/171 = \mathbf{43.9\%}$ |

Simpson's paradox

Causal inference perspective

3-point shot \tilde{y} : If shot goes in $\tilde{y} = 1$, if not $\tilde{y} = 0$

Treatment \tilde{t} : Player who shoots

From observed data

$$P(\tilde{y} = 1 \mid \tilde{t} = \text{Curry}) = 0.417$$

$$P(\tilde{y} = 1 \mid \tilde{t} = \text{Lee}) = 0.439$$

Potential outcomes

$\widetilde{\text{po}}_{\text{Curry}}$: Outcome if Curry shoots

$\widetilde{\text{po}}_{\text{Curry}} = 1$ shot made, $\widetilde{\text{po}}_{\text{Curry}} = 0$ shot missed

$\widetilde{\text{po}}_{\text{Lee}}$: Outcome if Lee shoots

$\widetilde{\text{po}}_{\text{Lee}} = 1$ shot made, $\widetilde{\text{po}}_{\text{Lee}} = 0$ shot missed

What we actually observe:











$$\tilde{y} := \begin{cases} \widetilde{\text{po}}_{\text{Curry}} & \text{if } \tilde{t} = \text{Curry} \\ \widetilde{\text{po}}_{\text{Lee}} & \text{if } \tilde{t} = \text{Lee} \end{cases}$$

Was Lee the better shooter?

$$P(\widetilde{\text{po}}_{\text{Lee}} = 1) > P(\widetilde{\text{po}}_{\text{Curry}} = 1)?$$

Challenge: We cannot observe them simultaneously!

Observed data

| Treatment \tilde{t} | Observed outcome \tilde{y} | Outcome if Curry $\widetilde{po}_{\text{Curry}}$ | Outcome if Lee $\widetilde{po}_{\text{Lee}}$ |
|--------------------------|---|---|---|
| Curry |  |  | ? |
| Curry |  |  | ? |
| Lee |  | ? |  |
| Lee |  | ? |  |
| Lee |  | ? |  |

? are counterfactuals

When does this hold?

$$P(\widetilde{\text{po}}_{\text{Curry}} = 1) \stackrel{=?}{=} P(\tilde{y} = 1 \mid \tilde{t} = \text{Curry}) = P(\widetilde{\text{po}}_{\text{Curry}} = 1 \mid \tilde{t} = \text{Curry})$$

$$P(\widetilde{\text{po}}_{\text{Lee}} = 1) \stackrel{=?}{=} P(\tilde{y} = 1 \mid \tilde{t} = \text{Lee}) = P(\widetilde{\text{po}}_{\text{Lee}} = 1 \mid \tilde{t} = \text{Lee})$$

True if \tilde{t} and $\widetilde{\text{po}}_{\text{Curry}} / \widetilde{\text{po}}_{\text{Lee}}$ are independent

Are they independent?

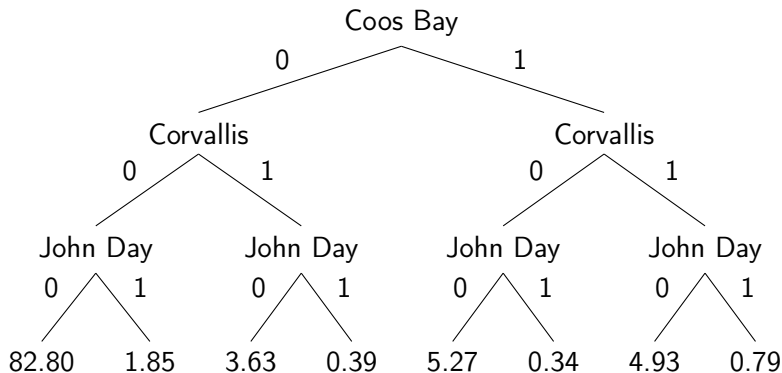
Shot distance

| | Stephen Curry | Courtney Lee |
|--------------------------------|--------------------|-------------------|
| Short threes (≤ 24 feet) | $45/90 = 50.0\%$ | $56/116 = 48.3\%$ |
| Long threes (> 24 feet) | $145/366 = 39.6\%$ | $19/55 = 34.5\%$ |

Distance is a **confounding factor**

We can adjust for it (if it's the only one!)

Toy examples have very few variables



Curse of dimensionality

Total weather stations in dataset: 134

Entries of joint pmf: $2^{134} \geq 10^{40}!!!$

Number of data: 8,760...

Dependencies **explode exponentially!**

Question 3

How to fight the curse of dimensionality

Make independence assumptions **even if they don't hold!**

Classification

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

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}

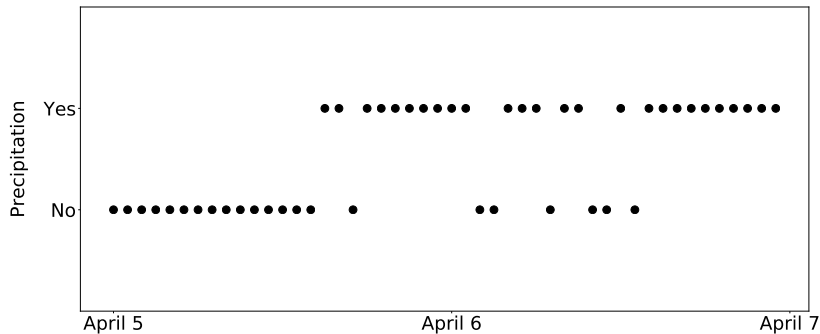
For new data vector x :

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

Problem: $p_{\tilde{y} | \tilde{x}}(\cdot | x)$ is impossible to estimate for all x !

Naive Bayes: Assume conditional independence given class

Time series



Data: x_1, x_2, \dots, x_n

x_i is measurement at time i

Modeling time series

Represent precipitation at each time by a random variable \tilde{a}_i

Then estimate joint pmf of $\tilde{a}_1, \dots, \tilde{a}_n$ from data

Entries in joint pmf? 2^n (if $n = 100$, more than 10^{30} !)

Curse of dimensionality

Markov property

$\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_n$ satisfy the Markov property if:

\tilde{a}_{i+1} is **conditionally independent** of $\tilde{a}_1, \dots, \tilde{a}_{i-1}$ given \tilde{a}_i

$$p_{\tilde{a}_{i+1} | \tilde{a}_1, \dots, \tilde{a}_i} (a_{i+1} | a_1, a_2, \dots, a_i) = p_{\tilde{a}_{i+1} | \tilde{a}_i} (a_{i+1} | a_i)$$

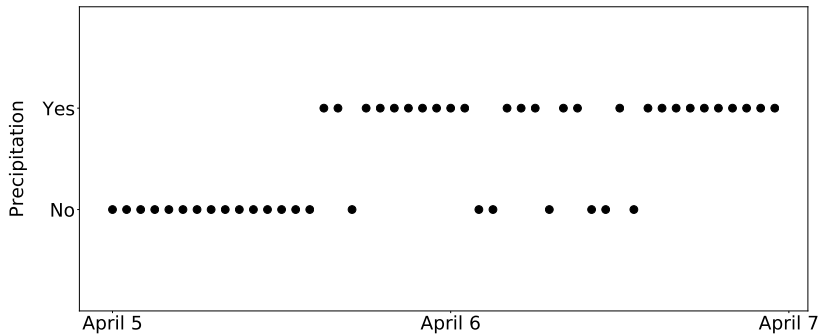
$$p_{\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_n} (a_1, a_2, \dots, a_n) = p_{\tilde{a}_1} (a_1) p_{\tilde{a}_2 | \tilde{a}_1} (a_2 | a_1) p_{\tilde{a}_3 | \tilde{a}_2} (a_3 | a_2) \dots$$

Time homogeneous finite state Markov chain

All transition probabilities are the same

$$p_{\tilde{a}_{i+1} | \tilde{a}_i} (a_{i+1} | a_i) = p_{\text{cond}} (a_{i+1} | a_i) \quad 1 \leq i \leq n - 1$$

Precipitation data



Precipitation data

Marginal probabilities

| | |
|------------|------|
| No | 88.7 |
| Yes | 11.3 |

1-step conditional probabilities

| | | |
|--------------------------------|----------------------------|------------|
| <i>Hour $h + 1$</i> | <i>Hour h</i> | |
| | No | Yes |
| | No | Yes |
| | 96.0 | 31.2 |
| | 4.0 | 68.8 |

What we have learned

How to jointly model multiple uncertain quantities

How to estimate causal relationships from data

How to fight the curse of dimensionality