Marginalization and Independence

Data Science: Jordan Boyd-Graber
University of Maryland
SLIDES ADAPTED FROM DAVE BLEI AND LAUREN HANNAH

Marginalization

If we know a joint distribution of multiple variables, what if we want to know the distribution of only one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\sum_y \sum_z P(X = x, Y = y, Z = z) = P(X)$$

Marginalization

If we know a joint distribution of multiple variables, what if we want to know the distribution of only one of the variables?

We can compute the distribution of $P(X)$ from $P(X, Y, Z)$ through *marginalization*:

$$\sum_y \sum_z P(X = x, Y = y, Z = z) = P(X)$$

We'll explain this notation more next week for now the formula is the most important part.

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather
- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15		

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny	
W=Cloudy	

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny	
W=Cloudy	

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny	.40
W=Cloudy	

Marginalization (from Leyton-Brown)

Joint distribution

temperature (T) and weather (W)

	T=Hot	T=Mild	T=Cold
W=Sunny	.10	.20	.10
W=Cloudy	.05	.35	.20

Marginalization allows us to compute distributions over smaller sets of variables:

- $P(X,Y) = \sum_z P(X,Y,Z=z)$
- Corresponds to summing out a table dimension
- New table still sums to 1

- Marginalize out weather

T=Hot	T=Mild	T=Cold
.15	.55	.30

- Marginalize out temperature

W=Sunny	.40
W=Cloudy	.60

Independence

Random variables X and Y are *independent* if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$.

Mathematical examples:

- If I flip a coin twice, is the second outcome independent from the first outcome?

Independence

Random variables X and Y are *independent* if and only if $P(X = x, Y = y) = P(X = x)P(Y = y)$.

Mathematical examples:

- If I flip a coin twice, is the second outcome independent from the first outcome?
- If I draw two socks from my (multicolored) laundry, is the color of the first sock independent from the color of the second sock?

Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Green Line is on schedule
  - snowfall in the Himalayas / your favorite color is blue

Independence

Intuitive Examples:

- Independent:
  - you use a Mac / the Green Line is on schedule
  - snowfall in the Himalayas / your favorite color is blue
- Not independent:
  - you vote for Larry Hogan / you are a Republican
  - there is a traffic jam Baltimore / there's a home game

Independence

Sometimes we make convenient assumptions.

- the values of two dice (ignoring gravity!)
- the value of the first die and the sum of the values
- whether it is raining and the number of taxi cabs
- whether it is raining and the amount of time it takes me to hail a cab
- the first two words in a sentence