Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Probability Distributions: Categorical and Poisson

Introduction to Data Science Algorithms

Dirk Grunwald

Categorical distribution

- The **categorical** distribution generalizes Bernoulli distribution over any number of values
  - Rolling a die
  - Selecting a card from a deck
- AKA *discrete* distribution.
  - Most general type of discrete distribution
  - specify all (but one) of the probabilities in the distribution

Categorical distribution

- If the categorical distribution is over K possible outcomes, then the distribution has K parameters.
- We will denote the parameters with a K-dimensional vector $\vec{\theta}$.
- The probability mass function can be written as:

$$f(x) = \prod_{k=1}^{K} \theta_k^{[x=k]}$$

  where the expression $[x = k]$ evaluates to 1 if the statement is true and 0 otherwise.
  - All this really says is that the probability of outcome x is equal to θ_x.
- The number of *free parameters* is $K-1$, since if you know $K-1$ of the parameters, the Kth parameter is constrained to sum to 1.

Categorical distribution

- Example: the roll of an unweighted die

$$
\begin{aligned}
P(X=1) &= \tfrac{1}{6} \\
P(X=2) &= \tfrac{1}{6} \\
P(X=3) &= \tfrac{1}{6} \\
P(X=4) &= \tfrac{1}{6} \\
P(X=5) &= \tfrac{1}{6} \\
P(X=6) &= \tfrac{1}{6}
\end{aligned}
$$

- If all outcomes have equal probability, this is called the *uniform* distribution.
- General notation: $P(X=x)=\theta_x$

Sampling from a categorical distribution

- How to randomly select a value distributed according to a categorical distribution?
- The idea is similar to randomly selected a Bernoulli-distributed value.
- Algorithm:
  1. Randomly generate a number between 0 and 1
     r = random(0, 1)
  2. For $k = 1, \dots, K$:
     - Return smallest r s.t. $r < \sum_{i=1}^{k} \theta_k$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667\\
P(X=2) &= \theta_2 &= 0.166667\\
P(X=3) &= \theta_3 &= 0.166667\\
P(X=4) &= \theta_4 &= 0.166667\\
P(X=5) &= \theta_5 &= 0.166667\\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667\\
P(X=2) &= \theta_2 &= 0.166667\\
P(X=3) &= \theta_3 &= 0.166667\\
P(X=4) &= \theta_4 &= 0.166667\\
P(X=5) &= \theta_5 &= 0.166667\\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?

- Return $X = 3$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01\\
P(X=2) &= \theta_2 &= 0.01\\
P(X=3) &= \theta_3 &= 0.01\\
P(X=4) &= \theta_4 &= 0.01\\
P(X=5) &= \theta_5 &= 0.01\\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$\begin{aligned} P(X=1) &= \theta_1 &= 0.01 \\ P(X=2) &= \theta_2 &= 0.01 \\ P(X=3) &= \theta_3 &= 0.01 \\ P(X=4) &= \theta_4 &= 0.01 \\ P(X=5) &= \theta_5 &= 0.01 \\ P(X=6) &= \theta_6 &= 0.95 \end{aligned}$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$\begin{aligned} P(X=1) &= \theta_1 &= 0.01 \\ P(X=2) &= \theta_2 &= 0.01 \\ P(X=3) &= \theta_3 &= 0.01 \\ P(X=4) &= \theta_4 &= 0.01 \\ P(X=5) &= \theta_5 &= 0.01 \\ P(X=6) &= \theta_6 &= 0.95 \end{aligned}$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$\begin{aligned} P(X=1) &= \theta_1 &= 0.01 \\ P(X=2) &= \theta_2 &= 0.01 \\ P(X=3) &= \theta_3 &= 0.01 \\ P(X=4) &= \theta_4 &= 0.01 \\ P(X=5) &= \theta_5 &= 0.01 \\ P(X=6) &= \theta_6 &= 0.95 \end{aligned}$$

Random number in $(0, 1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01\\
P(X=2) &= \theta_2 &= 0.01\\
P(X=3) &= \theta_3 &= 0.01\\
P(X=4) &= \theta_4 &= 0.01\\
P(X=5) &= \theta_5 &= 0.01\\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?
$r <$
$\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$\begin{aligned} P(X=1) &= \theta_1 &= 0.01 \\ P(X=2) &= \theta_2 &= 0.01 \\ P(X=3) &= \theta_3 &= 0.01 \\ P(X=4) &= \theta_4 &= 0.01 \\ P(X=5) &= \theta_5 &= 0.01 \\ P(X=6) &= \theta_6 &= 0.95 \end{aligned}$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?
$r <$
$\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6$?

- Return $X = 6$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01\\
P(X=2) &= \theta_2 &= 0.01\\
P(X=3) &= \theta_3 &= 0.01\\
P(X=4) &= \theta_4 &= 0.01\\
P(X=5) &= \theta_5 &= 0.01\\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$: $r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6$?

- Return $X = 6$

- We will always return $X = 6$ unless our random number $r < 0.05$.
  - 6 is the most probable outcome

Multinomial distribution

- Recall: the binomial distribution is the number of successes from multiple Bernoulli success/fail events
- The **multinomial** distribution is the number of different outcomes from multiple *categorical* events
    - It is a generalization of the binomial distribution to more than two possible outcomes
    - As with the binomial distribution, each categorical event is assumed to be independent
    - **Bernoulli : binomial :: categorical : multinomial**
- Examples:
    - The number of times each face of a die turned up after 50 rolls
    - The number of times each suit is drawn from a deck of cards after 10 draws

Multinomial distribution

- Notation: let $\vec{X}$ be a vector of length K, where X_k is a random variable that describes the number of times that the kth value was the outcome out of N categorical trials.
  - The possible values of each X_k are integers from 0 to N
  - All X_k values must sum to N: $\sum_{k=1}^{K} X_k = N$
- Example: if we roll a die 10 times, suppose it comes up with the following values:

$$\vec{X} = <1,0,3,2,1,3>$$

$$X_1 = 1$$
$$X_2 = 0$$
$$X_3 = 3$$
$$X_4 = 2$$
$$X_5 = 1$$
$$X_6 = 3$$

- The multinomial distribution is a *joint* distribution over multiple random variables: $P(X_1, X_2, \ldots, X_K)$

Multinomial distribution

- Suppose we roll a die 3 times. There are 216 (6^3) possible outcomes:

$$
\begin{aligned}
P(111) &= P(1)P(1)P(1) &= 0.00463 \\
P(112) &= P(1)P(1)P(2) &= 0.00463 \\
P(113) &= P(1)P(1)P(3) &= 0.00463 \\
P(114) &= P(1)P(1)P(4) &= 0.00463 \\
P(115) &= P(1)P(1)P(5) &= 0.00463 \\
P(116) &= P(1)P(1)P(6) &= 0.00463 \\
\ldots & \quad \ldots & \ldots \\
P(665) &= P(6)P(6)P(5) &= 0.00463 \\
P(666) &= P(6)P(6)P(6) &= 0.00463
\end{aligned}
$$

- What is the probability of a particular vector of counts after 3 rolls?

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} = < 0, 1, 0, 0, 2, 0 >$

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} = < 0, 1, 0, 0, 2, 0 >$
  - $P(\vec{X}) = P(255) + P(525) + P(552) = 0.01389$

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} =< 0, 1, 0, 0, 2, 0 >$
  - $P(\vec{X}) = P(255) + P(525) + P(552) = 0.01389$
- Example 2: $\vec{X} =< 0, 0, 1, 1, 1, 0 >$

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} = <0, 1, 0, 0, 2, 0>$
  - $P(\vec{X}) = P(255) + P(525) + P(552) = 0.01389$
- Example 2: $\vec{X} = <0, 0, 1, 1, 1, 0>$
  - $P(\vec{X}) = P(345) + P(354) + P(435) + P(453) + P(534) + P(543) = 0.02778$

Multinomial distribution

- The probability mass function for the multinomial distribution is:

$$f(\vec{x}) = \underbrace{\frac{N!}{\prod_{k=1}^{K} x_k!}}_{\text{Generalization of binomial coefficient}} \prod_{k=1}^{K} \theta_k^{x_k}$$

- Like categorical distribution, multinomial has a K-length parameter vector $\vec{\theta}$ encoding the probability of each outcome.
- Like binomial, the multinomial distribution has a additional parameter N, which is the number of events.

Multinomial distribution: summary

- Categorical distribution is multinomial when $N = 1$.
- Sampling from a multinomial: same code repeated N times.
  - Remember that each categorical trial is independent.
  - Question: Does this mean the count values (i.e., each X_1, X_2, etc.) are independent?

Multinomial distribution: summary

- Categorical distribution is multinomial when $N = 1$.
- Sampling from a multinomial: same code repeated N times.
  - Remember that each categorical trial is independent.
  - Question: Does this mean the count values (i.e., each X_1, X_2, etc.) are independent?
    - No! If $N = 3$ and $X_1 = 2$, then X_2 can be no larger than 1 (must sum to N).

Multinomial distribution: summary

- Categorical distribution is multinomial when $N = 1$.
- Sampling from a multinomial: same code repeated N times.
  - Remember that each categorical trial is independent.
  - Question: Does this mean the count values (i.e., each X_1, X_2, etc.) are independent?
    - No! If $N = 3$ and $X_1 = 2$, then X_2 can be no larger than 1 (must sum to N).

- Remember this analogy:
  - **Bernoulli : binomial :: categorical : multinomial**