COMP 761: Lecture 19 – Probability III

David Rolnick

October 19, 2020

Problem

What is the variance of the random variable that takes all values between a and b with equal probability? (uniform distribution)

(Please don't post your ideas in the chat just yet, we'll discuss the problem soon in class.)

Course Announcements

Course Announcements

- Office hours today right after class

Conditional prob and independence

Conditional prob and independence

- Good point raised in the Slack: If $p(A \mid B) = p(A)$, are A and B independent?

Conditional prob and independence

- Good point raised in the Slack: If $p(A \mid B) = p(A)$, are A and B independent?
- Yes!

$$p(A \mid B)p(B) = p(A \cap B),$$

so if $p(A \mid B) = p(A)$ then $p(A)p(B) = p(A \cap B)$.

Conditional prob and independence

- Good point raised in the Slack: If $p(A \mid B) = p(A)$, are A and B independent?
- Yes!

$$p(A \mid B)p(B) = p(A \cap B),$$

  so if $p(A \mid B) = p(A)$ then $p(A)p(B) = p(A \cap B)$.
- Reverse is true too: if A and B independent, then $p(A \mid B) = p(A)$ and $p(B \mid A) = p(B)$.

Intuition for entropy

Intuition for entropy

- How much information is needed to represent events from a probability distribution?

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.
- So that requires 2 bits of information.

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.
- So that requires 2 bits of information.
- If you had 8 events with probability $1/8$, it would be 3 bits.

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.
- So that requires 2 bits of information.
- If you had 8 events with probability $1/8$, it would be 3 bits.
- In general, for probability p, would be $\log_2 1/p = -\log_2 p$.

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.
- So that requires 2 bits of information.
- If you had 8 events with probability $1/8$, it would be 3 bits.
- In general, for probability p, would be $\log_2 1/p = -\log_2 p$.
- If 1 event has probability $1/2$ and there are 2 of probability $1/4$, then most efficient to use $0, 10, 11$ (essentially, merging 00 & 01).

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.
- So that requires 2 bits of information.
- If you had 8 events with probability $1/8$, it would be 3 bits.
- In general, for probability p, would be $\log_2 1/p = -\log_2 p$.
- If 1 event has probability $1/2$ and there are 2 of probability $1/4$, then most efficient to use $0, 10, 11$ (essentially, merging 00 & 01).
- Total expected number of bits: $(1/2) \cdot 1 + (1/4) \cdot 2 + (1/4) \cdot 2$.

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.
- So that requires 2 bits of information.
- If you had 8 events with probability $1/8$, it would be 3 bits.
- In general, for probability p, would be $\log_2 1/p = -\log_2 p$.
- If 1 event has probability $1/2$ and there are 2 of probability $1/4$, then most efficient to use $0, 10, 11$ (essentially, merging 00 & 01).
- Total expected number of bits: $(1/2) \cdot 1 + (1/4) \cdot 2 + (1/4) \cdot 2$.
- Another example, if probabilities $1/2, 1/4, 1/8, 1/8$:

$$(1/2) \cdot 1 + (1/4) \cdot 2 + (1/8) \cdot 3 + (1/8) \cdot 3.$$

Intuition for entropy

- How much information is needed to represent events from a probability distribution?
- If you have 4 events each with probability $1/4$, a compressed way to represent them is $00, 01, 10, 11$.
- So that requires 2 bits of information.
- If you had 8 events with probability $1/8$, it would be 3 bits.
- In general, for probability p, would be $\log_2 1/p = -\log_2 p$.
- If 1 event has probability $1/2$ and there are 2 of probability $1/4$, then most efficient to use $0, 10, 11$ (essentially, merging 00 & 01).
- Total expected number of bits: $(1/2) \cdot 1 + (1/4) \cdot 2 + (1/4) \cdot 2$.
- Another example, if probabilities $1/2, 1/4, 1/8, 1/8$:

$$(1/2) \cdot 1 + (1/4) \cdot 2 + (1/8) \cdot 3 + (1/8) \cdot 3.$$

- Generalization:

$$\sum_x p(X = x) \log_2(1/p(X = x)) = -\sum_x p(X = x) \log_2(p(X = x)).$$

Entropy

Entropy

- The *entropy* of a probability distribution is defined like this, only base e (easier to work with, just differs by a multiplicative constant):
$$H(p) = -\sum_x p(x)\log(p(x)).$$

Entropy

- The *entropy* of a probability distribution is defined like this, only base e (easier to work with, just differs by a multiplicative constant):
$$H(p) = -\sum_x p(x) \log(p(x)).$$
- (Using shorthand $p(x)$ for $p(X = x)$.)
- Entropy is always nonnegative - why?

Entropy

- The *entropy* of a probability distribution is defined like this, only base e (easier to work with, just differs by a multiplicative constant):
$$H(p) = -\sum_x p(x)\log(p(x)).$$
- (Using shorthand $p(x)$ for $p(X = x)$.)
- Entropy is always nonnegative - why?
- We have $p(x) \geq 0$ and $\log(p(x)) < 0$, so every term in the sum is negative.

Entropy

- The *entropy* of a probability distribution is defined like this, only base e (easier to work with, just differs by a multiplicative constant):
$$H(p) = -\sum_x p(x)\log(p(x)).$$
- (Using shorthand $p(x)$ for $p(X = x)$.)
- Entropy is always nonnegative - why?
- We have $p(x) \geq 0$ and $\log(p(x)) < 0$, so every term in the sum is negative.
- When is $H(p) = 0$?

Entropy

- The *entropy* of a probability distribution is defined like this, only base e (easier to work with, just differs by a multiplicative constant):
$$H(p) = -\sum_x p(x)\log(p(x)).$$
- (Using shorthand $p(x)$ for $p(X = x)$.)
- Entropy is always nonnegative - why?
- We have $p(x) \geq 0$ and $\log(p(x)) < 0$, so every term in the sum is negative.
- When is $H(p) = 0$?
- If the only event with nonzero probability has $p(x) = 1$, so zero uncertainty.

Entropy

- The *entropy* of a probability distribution is defined like this, only base e (easier to work with, just differs by a multiplicative constant):
$$H(p) = -\sum_x p(x) \log(p(x)).$$
- (Using shorthand $p(x)$ for $p(X = x)$.)
- Entropy is always nonnegative - why?
- We have $p(x) \geq 0$ and $\log(p(x)) < 0$, so every term in the sum is negative.
- When is $H(p) = 0$?
- If the only event with nonzero probability has $p(x) = 1$, so zero uncertainty.
- Continuous set of events:
$$H(p) = -\int p(x) \log(p(x))\, dx.$$

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different values.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different values.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- How do we get an upper bound for this?

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different values.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- How do we get an upper bound for this?
- How could we use Jensen's inequality? What function?

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different values.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- How do we get an upper bound for this?
- How could we use Jensen's inequality? What function?
- Let's try $z \log z$. Is it concave or convex?

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different values.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- How do we get an upper bound for this?
- How could we use Jensen's inequality? What function?
- Let's try $z \log z$. Is it concave or convex?
- We have

$$\frac{d}{dz}(z \log z) = z\left(\frac{1}{z}\right) + 1\,(\log z) = 1 + \log z$$

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different values.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- How do we get an upper bound for this?
- How could we use Jensen's inequality? What function?
- Let's try $z \log z$. Is it concave or convex?
- We have

$$\frac{d}{dz}(z \log z) = z\left(\frac{1}{z}\right) + 1\,(\log z) = 1 + \log z$$

$$\frac{d^2}{dz^2}(z \log z) = \frac{d}{dz}\,(1 + \log z) = \frac{1}{z}.$$

Entropy

If there are n different values for X, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different values.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- How do we get an upper bound for this?
- How could we use Jensen's inequality? What function?
- Let's try $z \log z$. Is it concave or convex?
- We have

$$\frac{d}{dz}(z \log z) = z\left(\frac{1}{z}\right) + 1\,(\log z) = 1 + \log z$$
$$\frac{d^2}{dz^2}(z \log z) = \frac{d}{dz}(1 + \log z) = \frac{1}{z}.$$

- Since $z > 0$ we have that $z \log z$ is convex.

Entropy

If there are n different events, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different events.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- We have that $z \log z$ is convex.

Entropy

If there are n different events, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different events.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- We have that $z \log z$ is convex.
- Then:

$$\frac{\sum_{k=1}^{n} p_k \log(p_k)}{n} \geq \left(\frac{p_1 + \cdots + p_n}{n}\right) \log\left(\frac{p_1 + \cdots + p_n}{n}\right)$$

Entropy

If there are n different events, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different events.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- We have that $z \log z$ is convex.
- Then:

$$\begin{aligned} \frac{\sum_{k=1}^{n} p_k \log(p_k)}{n} &\geq \left(\frac{p_1 + \cdots + p_n}{n}\right) \log\left(\frac{p_1 + \cdots + p_n}{n}\right) \\ &= \frac{1}{n} \log\left(\frac{1}{n}\right). \end{aligned}$$

Entropy

If there are n different events, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different events.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- We have that $z \log z$ is convex.
- Then:

$$\begin{aligned}\frac{\sum_{k=1}^{n} p_k \log(p_k)}{n} &\geq \left(\frac{p_1 + \cdots + p_n}{n}\right) \log\left(\frac{p_1 + \cdots + p_n}{n}\right) \\ &= \frac{1}{n} \log\left(\frac{1}{n}\right).\end{aligned}$$

- Multiplying by $-n$, get $H \leq -\log(1/n) = \log(n)$.

Entropy

If there are n different events, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different events.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- We have that $z \log z$ is convex.
- Then:

$$\frac{\sum_{k=1}^{n} p_k \log(p_k)}{n} \geq \left(\frac{p_1 + \cdots + p_n}{n}\right) \log\left(\frac{p_1 + \cdots + p_n}{n}\right) = \frac{1}{n} \log\left(\frac{1}{n}\right).$$

- Multiplying by $-n$, get $H \leq -\log(1/n) = \log(n)$.
- Is this value achievable? If so, when?

Entropy

If there are n different events, what is the maximum possible value for the entropy?

- Let $p_1, \ldots, p_n$ be the probabilities of the different events.

$$H = -\sum_{k=1}^{n} p_k \log(p_k).$$

- We have that $z \log z$ is convex.
- Then:

$$\begin{aligned} \frac{\sum_{k=1}^{n} p_k \log(p_k)}{n} &\geq \left(\frac{p_1 + \cdots + p_n}{n}\right) \log\left(\frac{p_1 + \cdots + p_n}{n}\right) \\ &= \frac{1}{n} \log\left(\frac{1}{n}\right). \end{aligned}$$

- Multiplying by $-n$, get $H \leq -\log(1/n) = \log(n)$.
- Is this value achievable? If so, when?
- Yes! Equality in Jensen's Inequality when all p_k equal.

Conditional entropy

Conditional entropy

- What happens if we know that $X = x$ but don't know $Y = y$?

Conditional entropy

- What happens if we know that $X = x$ but don't know $Y = y$?
- How much additional information is needed to get Y?

Conditional entropy

- What happens if we know that $X = x$ but don't know $Y = y$?
- How much additional information is needed to get Y?
- Need $-\log(p(Y = y \mid X = x))$.

Conditional entropy

- What happens if we know that $X = x$ but don't know $Y = y$?
- How much additional information is needed to get Y?
- Need $-\log(p(Y = y \mid X = x))$.
- Finding the expected value of that:

$$
\begin{aligned}
H(Y \mid X) &= \sum_{x,y} p(X = x, Y = y) \left(-\log(p(Y = y \mid X = x))\right) \\
&= -\sum_{x,y} p(X = x, Y = y) \log\left(\frac{p(X = x, Y = y)}{p(X = x)}\right).
\end{aligned}
$$

Conditional entropy

- What happens if we know that $X = x$ but don't know $Y = y$?
- How much additional information is needed to get Y?
- Need $-\log(p(Y = y \mid X = x))$.
- Finding the expected value of that:

$$
\begin{aligned}
H(Y \mid X) &= \sum_{x,y} p(X = x, Y = y)\,(-\log(p(Y = y \mid X = x))) \\
&= -\sum_{x,y} p(X = x, Y = y) \log\left(\frac{p(X = x, Y = y)}{p(X = x)}\right).
\end{aligned}
$$

- This is the called the *conditional entropy*.

Conditional entropy

- What happens if we know that $X = x$ but don't know $Y = y$?
- How much additional information is needed to get Y?
- Need $-\log(p(Y = y \mid X = x))$.
- Finding the expected value of that:

$$H(Y \mid X) = \sum_{x,y} p(X = x, Y = y)\,(-\log(p(Y = y \mid X = x)))$$
$$= -\sum_{x,y} p(X = x, Y = y) \log\left(\frac{p(X = x, Y = y)}{p(X = x)}\right).$$

- This is the called the *conditional entropy*.
- Compare to the standard entropy:

$$H(X) = -\sum_{x} p(X = x) \log(p(X = x)).$$

Conditional entropy

- What happens if we know that $X = x$ but don't know $Y = y$?
- How much additional information is needed to get Y?
- Need $-\log(p(Y = y \mid X = x))$.
- Finding the expected value of that:

$$H(Y \mid X) = \sum_{x,y} p(X = x, Y = y) \left(-\log(p(Y = y \mid X = x))\right)$$
$$= -\sum_{x,y} p(X = x, Y = y) \log\left(\frac{p(X = x, Y = y)}{p(X = x)}\right).$$

- This is the called the *conditional entropy*.
- Compare to the standard entropy:

$$H(X) = -\sum_{x} p(X = x) \log(p(X = x)).$$

- As with $H(X)$, can show $H(Y \mid X) \geq 0$.

Mutual information

Mutual information

- We define the *mutual information*:

$$I(X; Y) = H(Y) - H(Y \mid X).$$

Mutual information

- We define the *mutual information*:

$$I(X;Y) = H(Y) - H(Y \mid X).$$

- Let's expand it:

$$I(X;Y) = -\sum_y p(Y=y)\log(p(Y=y))$$
$$+\sum_{x,y} p(X=x, Y=y)\log\left(\frac{p(X=x,Y=y)}{p(X=x)}\right)$$

Mutual information

- We define the *mutual information*:

$$I(X;Y) = H(Y) - H(Y \mid X).$$

- Let's expand it:

$$\begin{aligned}
I(X;Y) &= -\sum_y p(Y=y)\log(p(Y=y)) \\
&\quad + \sum_{x,y} p(X=x, Y=y)\log\left(\frac{p(X=x,Y=y)}{p(X=x)}\right) \\
&= -\sum_{x,y} p(X=x, Y=y)\log(p(Y=y)) \\
&\quad + \sum_{x,y} p(X=x, Y=y)\log\left(\frac{p(X=x,Y=y)}{p(X=x)}\right) \\
&= \sum_{x,y} p(X=x, Y=y)\log\left(\frac{p(X=x,Y=y)}{p(X=x)p(Y=y)}\right)
\end{aligned}$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= \sum_{x,y} p(X=x, Y=y) \log \left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)} \right).
\end{aligned}
$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= \sum_{x,y} p(X=x, Y=y) \log \left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)} \right).
\end{aligned}
$$

- What can we learn from this?

Mutual information

$$I(X;Y) = H(Y) - H(Y \mid X)$$
$$= \sum_{x,y} p(X = x, Y = y) \log \left(\frac{p(X = x, Y = y)}{p(X = x)p(Y = y)} \right).$$

- What can we learn from this?
- If we switch X and Y, it's the same:

$$I(X;Y) = I(Y;X) = H(X) - H(X \mid Y).$$

Mutual information

$$I(X;Y) = H(Y) - H(Y \mid X)$$
$$= \sum_{x,y} p(X=x, Y=y) \log\left(\frac{p(X=x,Y=y)}{p(X=x)p(Y=y)}\right).$$

- What can we learn from this?
- If we switch X and Y, it's the same:

$$I(X;Y) = I(Y;X) = H(X) - H(X \mid Y).$$

- If X and Y are independent, then

$$\log\left(\frac{p(X=x,Y=y)}{p(X=x)p(Y=y)}\right) = \log 1 = 0.$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= \sum_{x,y} p(X=x, Y=y) \log \left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)} \right).
\end{aligned}
$$

- What can we learn from this?
- If we switch X and Y, it's the same:

$$I(X;Y) = I(Y;X) = H(X) - H(X \mid Y).$$

- If X and Y are independent, then

$$\log \left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)} \right) = \log 1 = 0.$$

- Can think of $I(X;Y)$ as being the information gained about Y by knowing X, or vice versa.

Mutual information

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= \sum_{x,y} p(X=x, Y=y) \log \left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)} \right).
\end{aligned}
$$

Mutual information

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= \sum_{x,y} p(X=x, Y=y) \log \left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)} \right).
\end{aligned}
$$

- Intuitively should be true that $I(X;Y) \geq 0$.

Mutual information

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= \sum_{x,y} p(X=x, Y=y) \log\left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)}\right).
\end{aligned}
$$

- Intuitively should be true that $I(X;Y) \geq 0$.
- Can show by Jensen's inequality.

Mutual information

$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= \sum_{x,y} p(X=x, Y=y) \log\left(\frac{p(X=x, Y=y)}{p(X=x)p(Y=y)}\right).
\end{aligned}
$$

- Intuitively should be true that $I(X;Y) \geq 0$.
- Can show by Jensen's inequality.
- Unfortunately it is **not** true that $\log\left(\frac{p(X=x,Y=y)}{p(X=x)p(Y=y)}\right)$ is always nonnegative.

Continuous probability distributions

Continuous probability distributions

- Suppose we have a continuous-valued variable X.

Continuous probability distributions

- Suppose we have a continuous-valued variable X.
- The probability of $X = x$ is always 0 – for example, essentially impossible that a random person is exactly 2.00000003 meters high.

Continuous probability distributions

- Suppose we have a continuous-valued variable X.
- The probability of $X = x$ is always 0 – for example, essentially impossible that a random person is exactly 2.00000003 meters high.
- In that case, we have a *probability density function* $p(x) \geq 0$.

Continuous probability distributions

- Suppose we have a continuous-valued variable X.
- The probability of $X = x$ is always 0 – for example, essentially impossible that a random person is exactly 2.00000003 meters high.
- In that case, we have a *probability density function* $p(x) \geq 0$.
- $p(x)$ is not the probability of $X = x$, because that would be 0.

Continuous probability distributions

- Suppose we have a continuous-valued variable X.
- The probability of $X = x$ is always 0 – for example, essentially impossible that a random person is exactly 2.00000003 meters high.
- In that case, we have a *probability density function* $p(x) \geq 0$.
- $p(x)$ is not the probability of $X = x$, because that would be 0.
- Instead, can talk about the probability that X is between a and b:

$$\int_a^b p(x)\, dx.$$

Continuous probability distributions

- Suppose we have a continuous-valued variable X.
- The probability of $X = x$ is always 0 – for example, essentially impossible that a random person is exactly 2.00000003 meters high.
- In that case, we have a *probability density function* $p(x) \geq 0$.
- $p(x)$ is not the probability of $X = x$, because that would be 0.
- Instead, can talk about the probability that X is between a and b:

$$\int_a^b p(x)\, dx.$$

- And we have

$$\int_{-\infty}^{\infty} p(x)\, dx = 1.$$

Continuous probability distributions

- Suppose we have a continuous-valued variable X.
- The probability of $X = x$ is always 0 – for example, essentially impossible that a random person is exactly 2.00000003 meters high.
- In that case, we have a *probability density function* $p(x) \geq 0$.
- $p(x)$ is not the probability of $X = x$, because that would be 0.
- Instead, can talk about the probability that X is between a and b:

$$\int_a^b p(x)\,dx.$$

- And we have

$$\int_{-\infty}^{\infty} p(x)\,dx = 1.$$

- Note that $p(x)$ can be bigger than 1 (though must be nonnegative or there would be an interval $[a, b]$ with negative probability).

Uniform distribution

Uniform distribution

- The *uniform distribution* on $[a, b]$ is the distribution where $a \leq x \leq b$ and $p(x)$ is equal everywhere.

Uniform distribution

- The *uniform distribution* on $[a, b]$ is the distribution where $a \leq x \leq b$ and $p(x)$ is equal everywhere.
- What is the right constant value C of $p(x)$?

Uniform distribution

- The *uniform distribution* on $[a, b]$ is the distribution where $a \leq x \leq b$ and $p(x)$ is equal everywhere.
- What is the right constant value C of $p(x)$?
- We want:

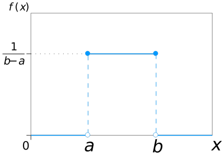
$$
\begin{aligned}
1 &= \int_{-\infty}^{\infty} p(x)\,dx = \int_a^b p(x)\,dx \\
&= \int_a^b C\,dx = (Cx)_{x=b} - (Cx)_{x=a} = C(b-a).
\end{aligned}
$$

Uniform distribution

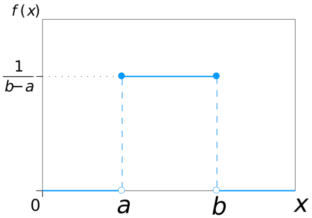
- The *uniform distribution* on $[a, b]$ is the distribution where $a \leq x \leq b$ and $p(x)$ is equal everywhere.
- What is the right constant value C of $p(x)$?
- We want:

$$1 = \int_{-\infty}^{\infty} p(x)\,dx = \int_{a}^{b} p(x)\,dx$$
$$= \int_{a}^{b} C\,dx = (Cx)_{x=b} - (Cx)_{x=a} = C(b-a).$$

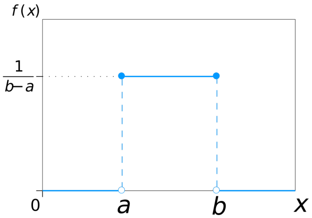
- So $p(x) = C = 1/(b-a)$.



Uniform distribution

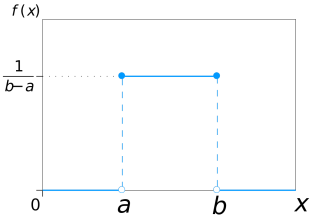


Uniform distribution



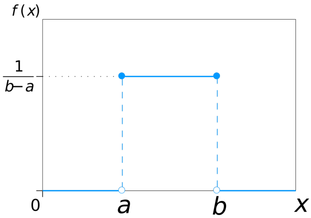
- $p(x) = 1/(b - a)$.

Uniform distribution



- $p(x) = 1/(b - a)$.
- What is the mean of $p(x)$?

Uniform distribution



- $p(x) = 1/(b-a)$.
- What is the mean of $p(x)$?
- Pretty clear that it is $(a+b)/2$, but can also calculate:

$$\mathbb{E}[X] = \int_a^b xp(x)\,dx = \int_a^b x/(b-a)dx$$
$$= \left(\frac{x^2}{2(b-a)}\right)_{x=b} - \left(\frac{x^2}{2(b-a)}\right)_{x=a} = \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

Uniform distribution

Uniform distribution

- $p(x) = 1/(b-a)$.

Uniform distribution

- $p(x) = 1/(b-a)$.
- What about variance?

Uniform distribution

- $p(x) = 1/(b-a)$.
- What about variance?

$$\mathbb{E}[X^2] = \int_a^b x^2 p(x)\, dx = \int_a^b x^2/(b-a) dx$$
$$= \left(\frac{x^3}{3(b-a)}\right)_{x=b} - \left(\frac{x^3}{3(b-a)}\right)_{x=a} = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}.$$

Uniform distribution

- $p(x) = 1/(b-a)$.
- What about variance?

$$\mathbb{E}[X^2] = \int_a^b x^2 p(x)\,dx = \int_a^b x^2/(b-a)dx$$
$$= \left(\frac{x^3}{3(b-a)}\right)_{x=b} - \left(\frac{x^3}{3(b-a)}\right)_{x=a} = \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}.$$

- So we have

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2$$
$$= \frac{4b^2+4ab+4a^2}{12} - \frac{3(a^2+2ab+b^2)}{12}$$
$$= \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}.$$

Univariate Gaussian

Univariate Gaussian

- The *Gaussian* or *normal* distribution $N(\mu, \sigma^2)$ is given by:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Univariate Gaussian

- The *Gaussian* or *normal* distribution $N(\mu, \sigma^2)$ is given by:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

- Can calculate (with some thorny integrals) that indeed $\int_{-\infty}^{\infty} p(x)\, dx = 1$

Univariate Gaussian

- The *Gaussian* or *normal* distribution $N(\mu, \sigma^2)$ is given by:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

- Can calculate (with some thorny integrals) that indeed $\int_{-\infty}^{\infty} p(x)\, dx = 1$
- And that the Gaussian has mean μ and variance σ^2.

Cool fact I: Adding independent Gaussian random variables

Cool fact I: Adding independent Gaussian random variables

- Suppose X and Y are independent random variables.

Cool fact I: Adding independent Gaussian random variables

- Suppose X and Y are independent random variables.
- We know $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ and $\text{Var}[X, Y] = \text{Var}[X] + \text{Var}[Y]$.

Cool fact I: Adding independent Gaussian random variables

- Suppose X and Y are independent random variables.
- We know $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ and $\text{Var}[X, Y] = \text{Var}[X] + \text{Var}[Y]$.
- If X, Y are Gaussian, it's better than that - their sum is Gaussian too.

Cool fact I: Adding independent Gaussian random variables

- Suppose X and Y are independent random variables.
- We know $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ and $\text{Var}[X, Y] = \text{Var}[X] + \text{Var}[Y]$.
- If X, Y are Gaussian, it's better than that - their sum is Gaussian too.
- If X is distributed by $N(\mu_1, \sigma_1^2)$ and Y is independently distributed by $N(\mu_2, \sigma_2^2)$, then $X + Y$ is distributed by $N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Cool fact II: Maximum entropy distribution

Cool fact II: Maximum entropy distribution

- $N(\mu, \sigma^2)$ is the distribution with maximum entropy out of all distributions with mean μ and variance σ^2.

Cool fact II: Maximum entropy distribution

- $N(\mu, \sigma^2)$ is the distribution with maximum entropy out of all distributions with mean μ and variance σ^2.
- (Can be proved with Lagrange multipliers.)

Cool fact II: Maximum entropy distribution

- $N(\mu, \sigma^2)$ is the distribution with maximum entropy out of all distributions with mean μ and variance σ^2.
- (Can be proved with Lagrange multipliers.)
- What is the maximum entropy distribution if you just assume it has mean μ?

Cool fact II: Maximum entropy distribution

- $N(\mu, \sigma^2)$ is the distribution with maximum entropy out of all distributions with mean μ and variance σ^2.
- (Can be proved with Lagrange multipliers.)
- What is the maximum entropy distribution if you just assume it has mean μ?
- Suppose $p(x) = 1/n$ for n different values with mean μ.

Cool fact II: Maximum entropy distribution

- $N(\mu, \sigma^2)$ is the distribution with maximum entropy out of all distributions with mean μ and variance σ^2.
- (Can be proved with Lagrange multipliers.)
- What is the maximum entropy distribution if you just assume it has mean μ?
- Suppose $p(x) = 1/n$ for n different values with mean μ.
- Then, entropy is $\log(n)$.

Cool fact II: Maximum entropy distribution

- $N(\mu, \sigma^2)$ is the distribution with maximum entropy out of all distributions with mean μ and variance σ^2.
- (Can be proved with Lagrange multipliers.)
- What is the maximum entropy distribution if you just assume it has mean μ?
- Suppose $p(x) = 1/n$ for n different values with mean μ.
- Then, entropy is $\log(n)$.
- By taking $n \to \infty$, we can make the entropy arbitrarily big.

Cool fact II: Maximum entropy distribution

- $N(\mu, \sigma^2)$ is the distribution with maximum entropy out of all distributions with mean μ and variance σ^2.
- (Can be proved with Lagrange multipliers.)
- What is the maximum entropy distribution if you just assume it has mean μ?
- Suppose $p(x) = 1/n$ for n different values with mean μ.
- Then, entropy is $\log(n)$.
- By taking $n \to \infty$, we can make the entropy arbitrarily big.
- So there isn't a maximum if don't constrain variance! Entropy can go to infinity.

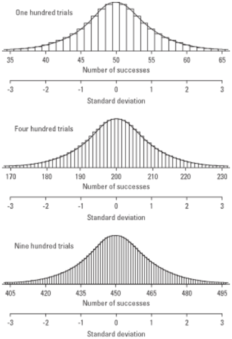
Cool fact III: Limit of binomial distribution

Cool fact III: Limit of binomial distribution

- The graph of $\binom{n}{k}$ for $k = 0, 1, \ldots, n$ approaches the Gaussian.

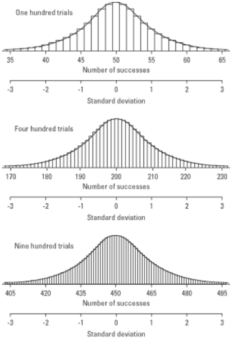
Cool fact III: Limit of binomial distribution

- The graph of $\binom{n}{k}$ for $k = 0, 1, \ldots, n$ approaches the Gaussian.
- Therefore, number of heads for n coin flips is approx Gaussian:



Cool fact III: Limit of binomial distribution

- The graph of $\binom{n}{k}$ for $k = 0, 1, \ldots, n$ approaches the Gaussian.
- Therefore, number of heads for n coin flips is approx Gaussian:



- More generally, the *Central Limit Theorem* says the sum of $X_1, \ldots, X_n$ drawn independently from the same distribution approaches a Gaussian as $n \to \infty$ regardless of the distribution.

Random matrices

Random matrices

- Random things can be surprisingly predictable.

Random matrices

- Random things can be surprisingly predictable.
- Given a large number n of fair coin flips, the expected number of heads is $n/2$.

Random matrices

- Random things can be surprisingly predictable.
- Given a large number n of fair coin flips, the expected number of heads is $n/2$.
- And we would be shocked to get $n/4$.

Random matrices

- Random things can be surprisingly predictable.
- Given a large number n of fair coin flips, the expected number of heads is $n/2$.
- And we would be shocked to get $n/4$.
- Same thing for big random matrices.

Random matrices

- Random things can be surprisingly predictable.
- Given a large number n of fair coin flips, the expected number of heads is $n/2$.
- And we would be shocked to get $n/4$.
- Same thing for big random matrices.
- Random matrix is an entire field :)

Random matrices

- Random things can be surprisingly predictable.
- Given a large number n of fair coin flips, the expected number of heads is $n/2$.
- And we would be shocked to get $n/4$.
- Same thing for big random matrices.
- Random matrix is an entire field :)
- Useful in everything from quantum to machine learning.

Example: Wigner's semicircle law

Example: Wigner's semicircle law

- We know the eigenvalues of a real symmetric matrix $A \in \mathbb{R}^{n \times n}$ are real.

Example: Wigner's semicircle law

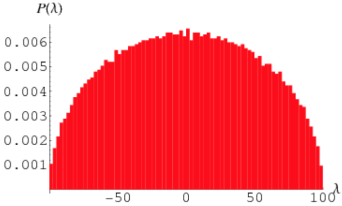
- We know the eigenvalues of a real symmetric matrix $A \in \mathbb{R}^{n \times n}$ are real.
- What are the eigenvalues if n goes to infinity, and A is random?

Example: Wigner's semicircle law

- We know the eigenvalues of a real symmetric matrix $A \in \mathbb{R}^{n \times n}$ are real.
- What are the eigenvalues if n goes to infinity, and A is random?
- Let's assume each entry taken independently of the others from a Gaussian.

Example: Wigner's semicircle law

- We know the eigenvalues of a real symmetric matrix $A \in \mathbb{R}^{n \times n}$ are real.
- What are the eigenvalues if n goes to infinity, and A is random?
- Let's assume each entry taken independently of the others from a Gaussian.
- Then if we do a histogram of n eigenvalues, it looks like a semicircle!



Next time!

Linear Programs I