Zaawansowane prawdopodobieństwo

W poprzedniej Części omówiliśmy podstawy prawdopodobieństwa i sposoby zastosowania prostych twierdzeń do złożonych zadań. Krótko mówiąc, prawdopodobieństwo to matematyka modelowania zdarzeń, które mogą wystąpić lub nie. Używamy formuł, aby opisać te zdarzenia, a nawet przyjrzeć się, jak wiele zdarzeń może zachowywać się razem. Tu przyjrzymy się bardziej skomplikowanym twierdzeniom dotyczącym prawdopodobieństwa i sposobom ich wykorzystania jako predykcyjnego. Zaawansowane tematy, takie jak twierdzenie Bayesa i zmienne losowe, dają początek powszechnym algorytmom uczenia maszynowego, takim jak algorytm Naïve Bayes (również omówiony). Skupimy się na niektórych bardziej zaawansowanych tematach teorii prawdopodobieństwa, w tym na następujących tematach:

• Wyczerpujące wydarzenia

• Twierdzenie Bayesa

• Podstawowe zasady przewidywania

• Zmienne losowe

Mamy jeszcze jedną definicję, na którą musimy się przyjrzeć, zanim zaczniemy (ostatnią przed zabawnymi rzeczami, obiecuję). Musimy przyjrzeć się zbiorowo wyczerpującym wydarzeniom.

Zbiorowo wyczerpujące wydarzenia

Gdy dany zestaw dwóch lub więcej zdarzeń musi zajść przynajmniej jedno ze zdarzeń, wówczas mówi się, że taki zestaw zdarzeń jest zbiorowo wyczerpujący. Rozważ następujące przykłady:

• Biorąc pod uwagę zestaw zdarzeń {temperatura < 60, temperatura > 90}, zdarzenia te nie są łącznie wyczerpujące, ponieważ istnieje trzecia opcja, która nie jest podana w tym zestawie zdarzeń: Temperatura może wynosić od 60 do 90. Jednak są wzajemnie wyczerpujące, ponieważ oba nie mogą się zdarzyć w tym samym czasie.

• W rzucie kostką zbiór wydarzeń związanych z rzutem {1, 2, 3, 4, 5 lub 6} jest łącznie wyczerpujący, ponieważ są to jedyne możliwe zdarzenia i przynajmniej jedno z nich musi się wydarzyć.

Ponowne przemyślenia bayesowskie

W ostatniej części rozmawialiśmy bardzo krótko o bayesowskich sposobach myślenia. Krótko mówiąc, mówiąc o Bayesie, mówisz o następujących trzech rzeczach i o tym, jak wszystkie ze sobą współdziałają:

• Wcześniejsza dystrybucja

• Dystrybucja tylna

• Prawdopodobieństwo

Zasadniczo zajmujemy się znalezieniem tyłu. To właśnie chcemy wiedzieć. Innym sposobem wyrażenia bayesowskiego sposobu myślenia jest to, że dane kształtują i aktualizują nasze przekonania. Mamy prawdopodobieństwo a priori, czyli to, co naiwnie myślimy o hipotezie, a potem mamy prawdopodobieństwo a posteriori, czyli to, co myślimy o hipotezie, biorąc pod uwagę pewne dane.

Twierdzenie Bayesa

Twierdzenie Bayesa jest dużym wynikiem wnioskowania bayesowskiego. Zobaczmy, jak to się w ogóle dzieje. Przypomnijmy, że wcześniej zdefiniowaliśmy:

• P(A) = Prawdopodobieństwo wystąpienia zdarzenia A

• P(A|B) = Prawdopodobieństwo wystąpienia A, zakładając, że wystąpiło B

• P(A, B) = Prawdopodobieństwo wystąpienia A i B

• P(A, B) = P(A) * P(B|A)

Ten ostatni punkt można odczytać jako prawdopodobieństwo wystąpienia A i B to prawdopodobieństwo wystąpienia A razy prawdopodobieństwo wystąpienia B, biorąc pod uwagę, że A już wystąpiło.To z tego ostatniego punktu nabiera kształtu twierdzenie Bayesa. Wiemy to:

P(A, B) = P(A) * P(B|A)

P(B, A) = P(B) * P(A|B)

P(A, B) = P(B, A)

Więc:

P(B) * P(A|B) = P(A) * P(B|A)

Dzieląc obie strony przez P(B) otrzymujemy twierdzenie Bayesa, jak pokazano:

P(A,B) = P(A) * P(B|A) / P(B)

Możesz myśleć o twierdzeniu Bayesa w następujący sposób:

• Jest to sposób na przejście z P(A|B) do P(B|A) (jeśli masz tylko jeden)

• Jest to sposób na uzyskanie P(A|B), jeśli już znasz P(A) (bez znajomości B)

Spróbujmy pomyśleć o Bayesie, używając terminów hipoteza i dane. Załóżmy, że H = twoja hipoteza na temat podanych danych, a D = dane, które otrzymujesz. Bayesa można interpretować jako próbę obliczenia P(H|D) (prawdopodobieństwo, że nasza hipoteza jest poprawna, biorąc pod uwagę dostępne dane). Aby skorzystać z naszej terminologii sprzed:

P(H|D) = P(D|H)P(H)/P(D)

• P(H) to prawdopodobieństwo hipotezy przed obserwowaniem danych, nazywane prawdopodobieństwem a priori lub po prostu przed

• P(H|D) to to, co chcemy obliczyć, prawdopodobieństwo hipotezy po zaobserwowaniu danych, zwane a posteriori

• P(D|H) to prawdopodobieństwo danych w ramach danej hipotezy, zwane prawdopodobieństwem

• P(D) to prawdopodobieństwo danych przy dowolnej hipotezie, zwane stałą normalizującą

Ta koncepcja nie odbiega od idei uczenia maszynowego i analityki predykcyjnej. W wielu przypadkach, rozważając analitykę predykcyjną, wykorzystujemy podane dane do przewidywania wyniku. Używając obecnej terminologii, H (naszą hipotezę) można uznać za nasz wynik, a P(H|D) (prawdopodobieństwo, że nasza hipoteza jest prawdziwa, biorąc pod uwagę nasze dane) to inny sposób powiedzenia: jaka jest szansa, że moja hipoteza jest poprawna , biorąc pod uwagę dane przede mną?. Rzućmy okiem na przykład, jak możemy wykorzystać formułę Bayesa w miejscu pracy. Weź pod uwagę, że masz dwie osoby odpowiedzialne za pisanie postów na blogu dla Twojej firmy - Lucy i Avinash. Z poprzednich występów podobało ci się 80% prac Lucy i tylko 50% prac Avinasha. Nowy wpis na blogu przychodzi na twoje biurko rano, ale autor nie jest wymieniony. Uwielbiasz ten artykuł. A+. Co to jest prawdopodobieństwo, że pochodzi od Avinasha? Każdy bloger bloguje w bardzo podobnym tempie. Zanim zwariujemy, zróbmy to, co zrobiłby każdy doświadczony matematyk (a teraz ty). Zapiszmy wszystkie nasze informacje, jak pokazano:

• H = hipoteza = blog pochodzi z Avinasha

• D = dane = podobał Ci się post na blogu

P(H|D) = szansa, że pochodzi od Avinasha, biorąc pod uwagę, że go pokochałeś

P(D|H) = szansa, że to pokochałeś, biorąc pod uwagę, że pochodzi od Avinasha

P(H) = szansa, że artykuł pochodził od Avinasha

P(D) = szansa, że pokochasz artykuł

Zauważ, że niektóre z tych zmiennych nie mają prawie żadnego sensu bez kontekstu. P(D), prawdopodobieństwo, że spodobałbyś się każdemu artykułowi postawionemu na biurku, jest dziwną koncepcją, ale uwierz mi, w kontekście formuły Bayesa, wkrótce będzie to istotne. Zauważ też, że w dwóch ostatnich punktach nie zakładają niczego innego. P(D) nie zakłada pochodzenia wpisu na blogu; pomyśl o P(D) tak, jakby artykuł został umieszczony na twoim biurku z jakiegoś nieznanego źródła, jaka jest szansa, że ci się spodoba? (ponownie, wiem, że to brzmi dziwnie poza kontekstem). Więc chcemy znać P(H|D). Spróbujmy użyć twierdzenia Bayesa, jak pokazano tutaj:

P(H|D) = P(D|H)P(H)/P(D)

Ale czy znamy liczby po prawej stronie tego równania? Twierdzę, że tak! Zobaczmy tutaj:

• P(H) to prawdopodobieństwo, że dany post na blogu pochodzi z Avinasha. Ponieważ blogerzy piszą w bardzo podobnym tempie, możemy założyć, że jest to 0,5, ponieważ mamy 50/50 szans, że pochodzi od któregokolwiek z blogerów (zauważ, że nie założyłem D, danych, w tym celu).

• P(D|H) to prawdopodobieństwo, że pokochasz post z Avinasha, które wcześniej powiedzieliśmy, że wynosi 50%, czyli 0,5.

• P(D) jest interesujące. To jest szansa, że ogólnie pokochasz artykuł. Oznacza to, że musimy wziąć pod uwagę scenariusz, jeśli post pochodził od Lucy czy Avinasha. Teraz, jeśli hipoteza tworzy zestaw, możemy użyć naszych praw prawdopodobieństwa, jak wspomniano w poprzednim rozdziale. Zestaw powstaje, gdy zestaw hipotez jest zarówno zbiorowo wyczerpujący, jak i wzajemnie się wykluczający. Mówiąc potocznie, w zestawie wydarzeń może wystąpić dokładnie jedna i tylko jedna hipoteza. W naszym przypadku dwie hipotezy są takie, że artykuł pochodzi od Lucy lub że artykuł pochodzi od Avinasha. Jest to zdecydowanie pakiet z następujących powodów:

- Przynajmniej jeden z nich to napisał

- Co najwyżej jeden z nich to napisał

- Dlatego napisał to dokładnie jeden z nich

Mając pakiet, możemy skorzystać z naszych reguł mnożenia i dodawania w następujący sposób:

$$D = (From\ Avinash\ AND\ loved\ it) \quad OR \quad (From\ Lucy\ AND\ loved\ it)$$

$$P(D) = P(Loved\ AND\ from\ Avinash) \quad OR \quad P(Loved\ AND\ from\ Lucy)$$

$$P(D) = P(From\ Avinash)P(Loved \mid form\ Avinash) \\ + P(from\ Lucy)P(Loved \mid from\ Lucy)$$

$$P(D) = .5(.5) + .5(.8) = \mathbf{.65}$$

Uff! Tak trzymać. Teraz możemy zakończyć nasze równanie, jak pokazano:

P(H|D) = P(D|H)P(H)/P(D)

P(H|D) = 5 * 0,5/0,65 = 0.38

Oznacza to, że istnieje 38% szans, że ten artykuł pochodzi od Avinasha. Co ciekawe, P(H) = 0,5 i P(H|D) = 0,38. Oznacza to, że bez jakichkolwiek danych szansa, że post na blogu pochodzi od Avinasha, była rzutem monetą, czyli 50/50. Biorąc pod uwagę pewne dane (Twoje przemyślenia na temat artykułu), zaktualizowaliśmy nasze przekonania dotyczące hipotezy i faktycznie obniżyliśmy szansę. O to właśnie chodzi w myśleniu bayesowskim — aktualizowanie naszych wcześniejszych przekonań na temat czegoś z wcześniejszego założenia, biorąc pod uwagę nowe dane na ten temat.

Więcej zastosowań twierdzenia Bayesa

Twierdzenie Bayesa pojawia się w wielu aplikacjach, zwykle wtedy, gdy musimy podejmować szybkie decyzje na podstawie danych i prawdopodobieństwa. Większość silników rekomendacji, takich jak Netflix, korzysta z niektórych elementów aktualizacji Bayesa. A jeśli zastanowisz się, dlaczego tak może być, ma to sens. Załóżmy, że w naszym uproszczonym świecie Netflix ma do wyboru tylko 10 kategorii. Załóżmy teraz, że przy braku danych szansa użytkownika na polubienie filmu komediowego z 10 kategorii wynosi 10% (tylko 1/10). Ok, teraz załóżmy, że użytkownik przyznał kilku filmom komediowym 5/5 gwiazdek. Teraz, gdy Netflix zastanawia się, jakie jest prawdopodobieństwo, że użytkownik polubi kolejną komedię, prawdopodobieństwo, że może polubić komedię, P(H|D), będzie większe niż losowe przypuszczenie wynoszące 10%! Wypróbujmy więcej przykładów zastosowania twierdzenia Bayesa przy użyciu większej ilości danych. Tym razem, zróbmy trochę bardziej twardzieli.

Przykład - Titanic

Bardzo znany zbiór danych obejmuje przyjrzenie się ocalałym z zatonięcia Titanica w 1912 roku. Użyjemy rachunku prawdopodobieństwa, aby dowiedzieć się, czy istnieją jakieś cechy demograficzne, które wykazywały związek z przeżyciem pasażerów. Przede wszystkim jesteśmy ciekawi, czy możemy wyizolować jakiekolwiek cechy naszego zbioru danych, które mogłyby nam powiedzieć więcej o typach ludzi, którzy prawdopodobnie przeżyli tę katastrofę. Najpierw przeczytajmy dane, jak pokazano tutaj:

```
titanic = pd.read_csv(data/titanic.csv')#read in a csv

titanic = titanic[['Sex', 'Survived']] #the Sex and Survived column

titanic.head()
```

	Sex	Survived
0	male	no
1	female	yes
2	female	yes
3	female	yes
4	male	no

W powyższej tabeli każdy wiersz reprezentuje jednego pasażera na statku i na razie przyglądamy się dwóm specyficznym cechom: płci osobnika i czy przeżył zatonięcie. Na przykład pierwszy wiersz reprezentuje mężczyznę, który nie przeżył, podczas gdy czwarty wiersz (z indeksem 3, pamiętaj, jak lista indeksów Pythona) reprezentuje kobietę, która przeżyła. Zacznijmy od podstaw. Zacznijmy od obliczenia prawdopodobieństwa, że jakakolwiek osoba na statku przeżyła, niezależnie od płci. Aby to zrobić, policzmy liczbę tak w kolumnie Przeżyli i podzielmy tę liczbę przez całkowitą liczbę wierszy, jak pokazano tutaj:

num_rows = float(titanic.shape[0]) # == 891 rows

p_survived = (titanic.Survived=="yes").sum() / num_rows # == .38

p_notsurvived = 1 - p_survived # == .61

Zauważ, że musiałem tylko obliczyć P (przeżył) i użyłem prawa sprzężonych prawdopodobieństw do obliczenia P (zmarł), ponieważ te dwa zdarzenia są komplementarne. Teraz obliczmy prawdopodobieństwo, że pojedynczy pasażer jest mężczyzną lub kobietą:

p_male = (titanic.Sex=="male").sum() / num_rows # == .65

p_female = 1 - p_male # == .35

Teraz zadajmy sobie pytanie, czy posiadanie określonej płci wpłynęło na przeżywalność? W tym celu możemy oszacować P (przeżył | kobieta) lub szansę, że ktoś przeżył, biorąc pod uwagę, że był kobietą. W tym celu musimy podzielić liczbę kobiet, które przeżyły, przez całkowitą liczbę kobiet, jak pokazano tutaj:

$$P(Survived \mid Female) = \frac{P(Female\ AND\ Survived)}{P(Female)}$$

number_of_women = titanic[titanic.Sex=='female'].shape[0] # == 314

women_who_lived = titanic[(titanic.Sex=='female') & (titanic.

Survived=='yes')].shape[0] # == 233

p_survived_given_woman = women_who_lived / float(number_of_women)

p_survived_given_woman # == .74

To całkiem duża różnica. Wydaje się, że płeć odgrywa dużą rolę w tym zbiorze danych.

Przykład - studia medyczne

Klasycznym zastosowaniem twierdzenia Bayesa jest interpretacja badań medycznych. Rutynowe testy na nielegalne zażywanie narkotyków są coraz powszechniejsze w miejscach pracy i szkołach. Firmy, które przeprowadzają te testy, utrzymują, że testy mają wysoką czułość, co oznacza, że prawdopodobnie dadzą pozytywny wynik, jeśli w ich systemie znajdują się leki. Twierdzą, że te testy są również bardzo specyficzne, co oznacza, że mogą dać wynik negatywny, jeśli nie ma leków. Średnio załóżmy, że czułość powszechnych testów narkotykowych wynosi około 60%, a swoistość około 99%. Oznacza to, że jeśli pracownik zażywa narkotyki, test ma 60% szans na wynik pozytywny, natomiast jeśli pracownik nie zażywa narkotyków, test ma 99% szans na wynik negatywny. Załóżmy teraz, że te testy są stosowane do siły roboczej, w której faktyczny wskaźnik używania narkotyków wynosi 5%. Prawdziwe pytanie dotyczy osób, które uzyskały pozytywny wynik testu, ilu faktycznie używa narkotyków? W kategoriach bayesowskich chcemy obliczyć prawdopodobieństwo zażywania narkotyków po pozytywnym teście.

Niech D = zdarzenie, w którym narkotyki są w użyciu

Niech E = zdarzenie, w którym test jest pozytywny

Niech N = przypadek, w którym narkotyki NIE są używane

Poszukujemy P(D|E).

Korzystając z twierdzenia Bayesa, możemy go ekstrapolować w następujący sposób:

P(D|E) = P(E}D)(PD)/P(E)

Poprzednie, P(D) to prawdopodobieństwo zażycia narkotyku, zanim zobaczymy wynik testu, który wynosi 5%. Prawdopodobieństwo, P(E|D), to prawdopodobieństwo pozytywnego wyniku testu przy założeniu zażywania narkotyków, które jest tym samym, co czułość testu. Stała normalizująca P(E) jest nieco trudniejsza. Musimy wziąć pod uwagę dwie rzeczy: P(E i D) oraz P(E i N). Zasadniczo musimy założyć, że test może być niepoprawny, gdy użytkownik nie używa narkotyków. Sprawdź następujące równania:

P(E) = P(E i D) lub P(E i N)

P(E) = P(D)P(E|D) + P(N)P(E|N)

P(E) = 0,5 * 0,6 + 0,95 * 0,1

P(E) = 0,0395

Tak więc nasze pierwotne równanie wygląda następująco:

P(D|E) = 0,6 * 0,5 / 0,0395

P(D}E) = 0,76

Oznacza to, że spośród osób, które uzyskały pozytywny wynik testu na zażywanie narkotyków, około jedna czwarta jest niewinna!

Zmienne losowe

Zmienna losowa wykorzystuje rzeczywiste wartości liczbowe do opisania zdarzenia probabilistycznego. W naszej poprzedniej pracy ze zmiennymi (zarówno matematycznymi, jak i programistycznymi) byliśmy przyzwyczajeni do tego, że zmienna przyjmuje określoną wartość. Na przykład, możemy mieć trójkąt, w którym otrzymamy zmienną h dla przeciwprostokątnej i musimy obliczyć długość przeciwprostokątnej. Możemy też mieć w Pythonie:

x = 5

Obie te zmienne są równe jednej wartości na raz. W zmiennej losowej podlegamy losowości, co oznacza, że wartości naszych zmiennych są, no właśnie, zmiennymi! W zależności od środowiska mogą przybierać różne wartości. Zmienna losowa nadal, jak pokazano wcześniej, przechowuje wartość. Główną różnicą między zmiennymi, jakie widzieliśmy, a zmienną losową jest fakt, że wartość zmiennej losowej może się zmieniać w zależności od sytuacji. Jeśli jednak zmienna losowa może mieć wiele wartości, w jaki sposób możemy je wszystkie śledzić? Każda wartość, jaką może przyjąć zmienna losowa, jest powiązana z wartością procentową. Dla każdej wartości, jaką może przyjąć zmienna losowa, istnieje jedno prawdopodobieństwo, że zmienna będzie tą wartością. Za pomocą zmiennej losowej możemy również uzyskać nasz rozkład prawdopodobieństwa losowej zmiennej, która podaje możliwe wartości zmiennej i ich prawdopodobieństwa. Napisane, zwykle używamy pojedynczych wielkich liter (głównie konkretnej litery X) do oznaczenia zmiennych losowych. Na przykład możemy mieć:

• X = wynik rzutu kostką

• Y = przychód osiągnięty przez firmę w tym roku

• Z = wynik kandydata w quizie z kodowania rozmowy kwalifikacyjnej (0-100%)

W efekcie zmienna losowa to funkcja, która odwzorowuje wartości z przestrzeni próbki zdarzenia (zbiór wszystkich możliwych wyników) na wartość prawdopodobieństwa (między 0 a 1). Pomyśl o wydarzeniu jako wyrażonym w następujący sposób:

f(zdarzenie) = prawdopodobieństwo

Przypisze prawdopodobieństwo każdej indywidualnej opcji. Istnieją dwa główne typy zmiennych losowych: dyskretne i ciągłe.

Dyskretne zmienne losowe

Dyskretna zmienna losowa przyjmuje tylko policzalną liczbę możliwych wartości. Na przykład wynik rzutu kostką, jak pokazano tutaj:

X = wynik pojedynczego rzutu kostką

Value	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$	$X=6$
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Zwróć uwagę, jak używam dużej litery X, aby zdefiniować zmienną losową. To powszechna praktyka. Zwróć także uwagę, w jaki sposób zmienna losowa odwzorowuje prawdopodobieństwo dla każdego indywidualnego wyniku. Zmienne losowe mają wiele właściwości, z których dwie to ich wartość oczekiwana i wariancja. Użyjemy funkcji masy prawdopodobieństwa (PMF) do opisania dyskretnej zmiennej losowej. Przybierają wygląd:

P(X = x) = PMF

Tak więc, dla rzutu kostką, P(X = 1) = 1/6 i P(X = 5) = 1/6.

Rozważ następujące przykłady zmiennych dyskretnych:

• Prawdopodobny wynik pytania ankietowego (na przykład w skali 1-10)

• Czy dyrektor generalny zrezygnuje w ciągu roku (prawda lub fałsz)

Oczekiwana wartość zmiennej losowej określa średnią wartość długiego przebiegu powtarzanych próbek zmiennej losowej. Jest to czasami nazywane średnią zmiennej. Na przykład zapoznaj się z następującym kodem Pythona, który definiuje zmienną losową rzutu kostką:

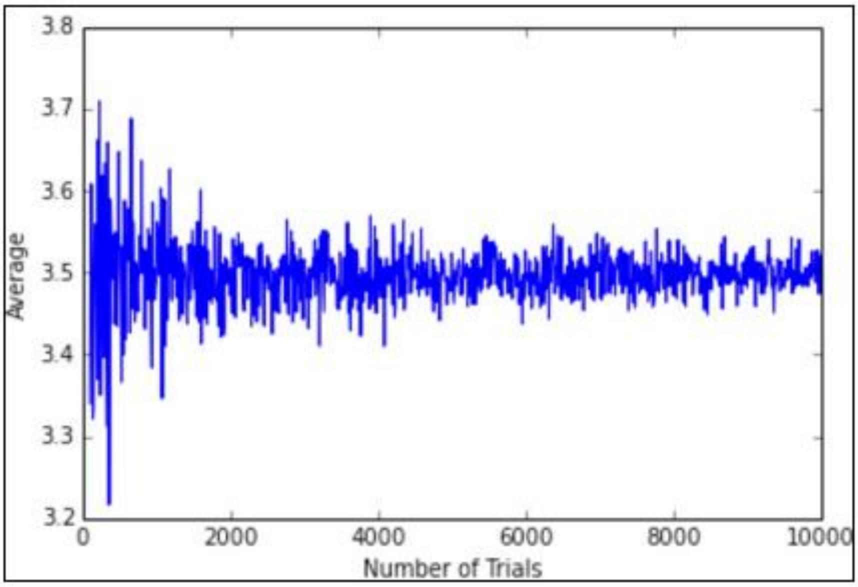
```
import random

def random_variable_of_dice_roll():

return random.randint(1, 7) # a range of (1,7) # includes 1, 2, 3,

4, 5, 6, but NOT 7
```

Ta funkcja wywoła zmienną losową i wyjdzie z odpowiedzią. Rzućmy 100 kostkami i uśrednimy wynik w następujący sposób:

```
trials = []

num_trials = 100

for trial in range(num_trials):

trials.append( random_variable_of_dice_roll() )

print sum(trials)/float(num_trials) # == 3.77
```

Tak więc wzięcie 100 rzutów kostką i uśrednienie ich daje nam wartość 3,77! Wypróbujmy to z szeroką gamą numerów próbnych, jak pokazano tutaj:

```
num_trials = range(100,10000, 10)

avgs = []

for num_trial in num_trials:

trials = []

for trial in range(1,num_trial):

trials.append( random_variable_of_dice_roll() )

avgs.append(sum(trials)/float(num_trial))

plt.plot(num_trials, avgs)

plt.xlabel('Number of Trials')

plt.ylabel("Average")
```



Powyższy wykres przedstawia średni rzut kostką, gdy przyglądamy się coraz większej liczbie rzutów kośćmi. Widzimy, że średni rzut kostką szybko zbliża się do 3,5. Jeśli spojrzymy w lewo na wykresie, zobaczymy, że jeśli rzucimy kostką tylko około 100 razy, nie gwarantujemy, że przeciętny rzut kostką wyniesie 3,5. Jednakże, jeśli rzucamy 10,000 kostkami jedna po drugiej, widzimy, że bardzo prawdopodobne jest, że średni rzut kostką wyniesie około 3,5. Dla dyskretnej zmiennej losowej możemy również użyć prostego wzoru, pokazanego poniżej, aby obliczyć wartość oczekiwaną:

Wartość oczekiwana = $E[X] = \mu X = \Sigma x_i p_i$

Gdzie x_i jest i-tym wynikiem, a p_i jest i-tym prawdopodobieństwem. Tak więc dla naszego rzutu kostką możemy znaleźć dokładną oczekiwaną wartość w następujący sposób:

$$\frac{1}{6}(1)+\frac{1}{6}(2)+\frac{1}{6}(3)+\frac{1}{6}(4)+\frac{1}{6}(5)+\frac{1}{6}(6)=3.5$$

Powyższy wynik pokazuje nam, że dla każdego rzutu kostką możemy „oczekiwać” rzutu kostką o wartości 3,5. Oczywiście nie ma to sensu, ponieważ nie możemy uzyskać 3,5 na rzucie kostką, ale ma to sens w kontekście wielu rzutów kośćmi. Jeśli rzucisz 10 000 kostkami, średni rzut powinien zbliżyć się do 3,5, jak pokazano na powyższym wykresie i kodzie. Średnia wartości oczekiwanej zmiennej losowej na ogół nie wystarcza, aby zrozumieć pełną ideę zmiennej. Z tego powodu wprowadzamy nową koncepcję, zwaną wariancją. Wariancja zmiennej losowej reprezentuje rozrzut zmiennej. Określa ilościowo zmienność wartości oczekiwanej. Wzór na wariancję dyskretnej zmiennej losowej wyraża się następująco:

$$\text{Variance} = V[X] = \sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$$

Gdzie x_i i p_i reprezentują te same wartości co poprzednio i reprezentuje oczekiwaną wartość zmiennej. W tym wzorze wspomniałem również o sigma X. Sigma, w tym przypadku, to odchylenie standardowe, które definiuje się po prostu jako pierwiastek kwadratowy z wariancji. Spójrzmy na bardziej skomplikowany przykład dyskretnej zmiennej losowej. Wariancję można traktować jako miernik dawania lub brania. Jeśli powiem, że możesz spodziewać się wygrania 100 $ z rozdania w pokera, możesz być bardzo szczęśliwy. Jeśli dołączę do tego stwierdzenia dodatkowy szczegół, że możesz wygrać 100 USD, dać lub wziąć 80 USD, masz teraz do czynienia z wieloma oczekiwaniami, co może być frustrujące i może sprawić, że gracz niechętny ryzyku będzie bardziej ostrożny przed dołączeniem do gry . Zwykle możemy powiedzieć, że mamy wartość oczekiwaną, podając lub przyjmując odchylenie standardowe. Weź pod uwagę, że Twój zespół mierzy sukces nowego produktu w skali Likerta, czyli w jednej z pięciu kategorii, gdzie wartość 0 oznacza całkowitą porażkę, a 4 oznacza wielki sukces. Szacują, że nowy projekt ma następujące szanse powodzenia na podstawie testów użytkowników i wstępnych wyników działania produktu. Najpierw musimy zdefiniować naszą zmienną losową. Niech zmienna losowa X reprezentuje sukces naszego produktu. X jest rzeczywiście dyskretną zmienną losową, ponieważ zmienna X może przyjąć tylko jedną z pięciu opcji: 0, 1, 2, 3 lub 4. Poniżej znajduje się rozkład prawdopodobieństwa naszej zmiennej losowej X. Zauważ, że mamy kolumnę dla każdego potencjalnego wyniku X, a poniżej każdego wyniku mamy prawdopodobieństwo, że ten konkretny wynik zostanie osiągnięty:

Value	$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$
Probability	0.02	0.07	0.25	0.4	0.26

Na przykład projekt ma 2% szansy na całkowitą porażkę i 26% szansy na wielki sukces! Naszą oczekiwaną wartość możemy obliczyć w następujący sposób:

E[X] = 0(0,02) + 1(0,07) + 2(0,25) + 3(0,4) + 4(0,26) = 2,81

Ta liczba oznacza, że menedżer może liczyć na sukces około 2,81 z tego projektu. Sama liczba ta nie jest zbyt użyteczna. Być może, mając do wyboru kilka produktów, oczekiwana wartość może być sposobem na porównanie potencjalnych sukcesów kilku produktów. Jednak w tym przypadku, gdy mamy do oceny tylko jeden produkt, będziemy potrzebować więcej. Sprawdźmy teraz wariancję, jak pokazano tutaj:

Wariancja=V[X]=σX2 = (x_i −μX)2pi = (0 − 2,81)2(0,02) + (1 − 2,81)2(0,07)+ (2 − 2,81)2(0,25) + (3 − 2,81)2(0,4) + (4 −2,81)2(0,26) = 0,93

Teraz, gdy mamy już zarówno odchylenie standardowe, jak i oczekiwaną wartość punktacji projektu, spróbujmy podsumować nasze wyniki. Można powiedzieć, że nasz projekt będzie miał oczekiwany wynik 2,81 plus minus 0,96, co oznacza, że można spodziewać się czegoś między 1,85 a 3,77. Tak więc jednym ze sposobów, w jaki możemy zająć się tym projektem, jest to, że prawdopodobnie będzie miał on wskaźnik sukcesu wynoszący 2,81, co daje lub bierze około punktu. Możesz pomyśleć, wow, Sinan, więc w najlepszym przypadku projekt będzie miał 3,8, a w najgorszym 1,8?. Nie do końca. Może być lepszy niż 4 i może być też gorszy niż 1,8. Aby zrobić ten jeden krok dalej obliczmy, co następuje:

P(X >= 3)

Najpierw poświęć chwilę i przekonaj się, że możesz przeczytać sobie tę formułę. O co pytam, gdy proszę o P(X >= 3)? Szczerze, poświęć chwilę i zastanów się. P(X >= 3) to prawdopodobieństwo, że nasza zmienna losowa przyjmie wartość co najmniej tak dużą jak 3. Innymi słowy, jaka jest szansa, że nasz produkt będzie miał ocenę sukcesu 3 lub wyższą? Aby to obliczyć, możemy obliczyć:

P(X >= 3) = P(X = 3) + P(X = 4) = 0,66 = 66%

Oznacza to, że mamy 66% szans, że nasz produkt oceni się na 3 lub 4. Innym sposobem obliczenia tego będzie sposób sprzężony, jak pokazano tutaj:

$P(X >= 3) = 1 – P(X < 3)$

Ponownie poświęć chwilę, aby przekonać się, że ta formuła się sprawdza. Twierdzę, że znalezienie prawdopodobieństwa, że produkt otrzyma ocenę co najmniej 3, jest takie samo jak 1 minus prawdopodobieństwo, że produkt otrzyma ocenę poniżej 3. Jeśli to prawda, to dwa zdarzenia ($X >=3$ a $X < 3$) muszą się wzajemnie uzupełniać. To oczywiście prawda! Produkt może mieć jedną z dwóch następujących opcji:

• mieć ocenę 3 lub wyższą

• Mieć ocenę poniżej 3

Sprawdźmy naszą matematykę:

$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = 0{,}02 + 0{,}07 + 0{,}25 = 0{,}034$

$1 – P(X < 3) = 1 - 0{,}34 = 0{,}66 = P(x >= 3)$

To się sprawdza!

Rodzaje dyskretnych zmiennych losowych

Możemy lepiej zrozumieć, jak działają zmienne losowe w praktyce, przyglądając się określonym typom zmiennych losowych. Te specyficzne typy zmiennych losowych modelują różne rodzaje sytuacji i kończą się ujawnieniem znacznie prostszych obliczeń dla bardzo złożonego modelowania zdarzeń.

Zmienne losowe dwumianowe

Pierwszy typ dyskretnej zmiennej losowej, któremu się przyjrzymy, nazywa się zmienną losową dwumianową. W przypadku dwumianowej zmiennej losowej patrzymy na ustawienie, w którym pojedyncze zdarzenie zdarza się w kółko i próbujemy policzyć, ile razy wynik jest dodatni. Zanim zrozumiemy samą zmienną losową, musimy przyjrzeć się warunkom, w jakich jest ona nawet odpowiednia.

Ustawienie dwumianowe ma następujące cztery warunki:

• Możliwe wyniki to sukces lub porażka

• Wyniki badań nie mogą wpływać na wyniki innego badania

• Ustalono liczbę prób (stała wielkość próby)

• Szansa powodzenia każdej próby musi zawsze wynosić p

Zmienna losowa dwumianowa to dyskretna zmienna losowa X, która zlicza liczbę sukcesów w układzie dwumianowym. Parametry to n = liczba prób i p = szansa powodzenia każdej próby.

Przykład - spotkania fundraisingowe:

Start-up bierze 20 spotkań VC, aby sfinansować i zliczyć otrzymywane oferty. Funkcja masy prawdopodobieństwa (PMF) dla dwumianowej zmiennej losowej jest następująca:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Tu

$$\binom{n}{k} = the\,binomial\,coefficient = \frac{n!}{(n-k)!k!}$$

Przykład - otwarcia restauracji

Nowa restauracja w mieście ma 20% szans na przetrwanie pierwszego roku. Jeśli w tym roku zostanie otwartych 14 restauracji, sprawdź prawdopodobieństwo, że dokładnie cztery restauracje przetrwają pierwszy rok otwarcia dla publiczności. Najpierw powinniśmy udowodnić, że jest to ustawienie dwumianowe:

• Możliwe wyniki to sukces lub porażka (restauracje albo przetrwają, albo nie)

• Wyniki prób nie mogą wpływać na wyniki innej próby (załóżmy, że otwarcie jednej restauracji nie wpływa na otwarcie i przetrwanie innej restauracji)

• Ustalono liczbę prób (otworzono 14 restauracji)

• Szansa powodzenia każdej próby musi zawsze wynosić p (zakładamy, że zawsze wynosi 20%)

Tutaj mamy nasze dwa parametry n = 14 i p = .2. Możemy więc teraz wstawić te liczby do naszego wzoru dwumianowego, jak pokazano tutaj:

$$P(X=4) = \binom{14}{4}.2^4.8^{10} = .17$$

Mamy więc 17% szans, że dokładnie 4 z tych restauracji będą otwarte za rok

Przykład - grupy krwi

Para ma 25% szans na urodzenie dziecka z grupą krwi O. Jaka jest szansa, że 3 z ich 5 dzieci ma krew typu O?

Niech X = liczba dzieci z krwią typu O z n = 5 i p = 0,25, jak pokazano tutaj:

P(X = 3) = 5 0,253 (0,75)5−3 = 10(0,25)3(0,75)2 = 0,087

Możemy obliczyć to prawdopodobieństwo dla wartości 0, 1, 2, 3, 4 i 5, aby uzyskać sens rozkładu prawdopodobieństwa:

value x_i	0	1	2	3	4	5
Probability	0.23730	0.39551	0.26367	0.08789	0.01465	0.00098

Stąd możemy obliczyć wartość oczekiwaną i wariancję tej zmiennej:

$$\text{Expected value} = E[X] = \mu_X = \sum x_i p_i = 1.25$$

$$\text{Variance} = V[X] = \sigma_X^2 = \sum (x_i - \mu_X)^2 p_i = 0.9375$$

Tak więc ta rodzina może spodziewać się prawdopodobnie 1 lub 2 dzieci z grupą krwi O! A jeśli chcemy poznać prawdopodobieństwo, że przynajmniej troje ich dzieci ma grupę krwi 0? Aby poznać prawdopodobieństwo, że co najmniej troje ich dzieci ma krew typu 0, możemy użyć następującego wzoru na dyskretne zmienne losowe:

$$P(X \geq 3) = P(X = 5) + P(X = 4) + P(X = 3)$$

$$= .00098 + .01465 + .08789 = 0.103$$

Tak więc istnieje około 10% szans, że troje ich dzieci ma grupę 0 krwi.

Skróty do dwumianowej wartości oczekiwanej i wariancji

Zmienne losowe dwumianowe mają specjalne obliczenia dla dokładnych wartości oczekiwanych wartości i wariancji. Jeśli X jest dwumianową zmienną losową, to:

E(X) = np

V(X) = np(1 − p)

W naszym poprzednim przykładzie możemy użyć następujących formuł, aby obliczyć dokładną wartość oczekiwaną i wariancję:

• E(X) = 0,25(5) = 1,25

• V(X) = 1,25(,75) = 0,9375

Zmienna losowa dwumianowa to dyskretna zmienna losowa, która zlicza liczbę sukcesów w układzie dwumianowym. Jest używany w wielu różnych eksperymentach opartych na danych, takich jak liczenie osób, które zarejestrują się w witrynie, mając szansę na konwersję, lub nawet, na prostym poziomie, przewidywanie ruchów cen akcji przy szansie na spadek (nie martw się; będziemy później stosować znacznie bardziej wyrafinowane modele do przewidywania rynku akcji).

Geometryczne zmienne losowe

Druga dyskretna zmienna losowa, której się przyjrzymy, nazywa się geometryczną zmienną losową. W rzeczywistości jest dość podobna do dwumianowej zmiennej losowej, ponieważ zajmujemy się ustawieniem, w którym pojedyncze zdarzenie ma miejsce w kółko. Jednak w przypadku ustawienia geometrycznego główną różnicą jest to, że nie ustalamy wielkości próbki. Nie idziemy na dokładnie 20 spotkań VC jako start-up, ani nie mamy dokładnie 5 dzieci. Zamiast tego w ustawieniu geometrycznym modelujemy liczbę prób, które będziemy musieli zobaczyć, zanim osiągniemy choćby jeden sukces. W szczególności ustawienie geometryczne ma następujące cztery warunki:

• Możliwe wyniki to sukces lub porażka

• Wyniki badań nie mogą wpływać na wyniki innego badania

• Nie ustalono liczby prób

• Szansa powodzenia każdej próby musi zawsze wynosić p

Zauważ, że są to dokładnie te same warunki, co zmienna dwumianowa, z wyjątkiem trzeciego warunku. Geometryczna zmienna losowa to dyskretna zmienna losowa X, która zlicza liczbę prób potrzebną do osiągnięcia jednego sukcesu. Parametry to p = szansa powodzenia każdej próby i (1 − p) = szansa niepowodzenia każdej próby. Aby przekształcić poprzednie przykłady dwumianowe w przykłady geometryczne, możemy wykonać następujące czynności:

• Policz liczbę spotkań VC, które musi odbyć start-up, aby uzyskać pierwsze tak

• Policz liczbę rzutów monetą potrzebną do uzyskania orła (tak, wiem, że to nudne, ale to dobry przykład!)

Wzór na PMF jest następujący:

$P(X = x) = (1-p)[x-1]p$

Zarówno ustawienia dwumianowe, jak i geometryczne wiążą się z wynikami, które są albo sukcesami, albo porażkami. Duża różnica polega na tym, że dwumianowe zmienne losowe mają ustaloną liczbę prób, oznaczoną jako n. Geometryczne zmienne losowe nie mają stałej liczby prób. Zamiast tego geometryczne zmienne losowe modelują liczbę próbek potrzebnych do uzyskania pierwszej udanej próby, niezależnie od tego, jaki sukces może oznaczać w tych warunkach eksperymentalnych.

Przykład – pogoda

Jest 34% szans na deszcz w dowolnym dniu kwietnia. Znajdź prawdopodobieństwo, że pierwszy dzień deszczu w kwietniu nastąpi czwartego kwietnia.

Niech X = liczba dni do deszczu (sukces) przy $p = 0{,}34$ i $(1 - p) = 0{,}66$

Tak więc $P(X = 8) = (0{,}66)8-1\ (0{,}34)$

$= (0{,}66)7\ (0{,}34)$

$= 0{,}01855$

Prawdopodobieństwo, że do czwartego kwietnia będzie padać, przedstawia się następująco:

$$P(X \leq 4) = P(1) + P(2) + P(3) + P(4) =$$

$$= .34 + .22 + .14 + .1 = .8$$

Tak więc istnieje 80% szans na to, że pierwszy deszcz miesiąca wystąpi w ciągu pierwszych czterech dni.

Skróty do geometrycznej wartości oczekiwanej i wariancji

Geometryczne zmienne losowe mają również specjalne obliczenia dla dokładnych wartości oczekiwanych wartości i wariancji. Jeśli X jest geometryczną zmienną losową, to

$E(X) = 1/p$

$V(X) = (1-p)/p2$

str 124

Zmienna losowa Poissona

Trzecim i ostatnim konkretnym przykładem dyskretnej zmiennej losowej jest zmienna losowa Poissona. Aby zrozumieć, dlaczego potrzebujemy tej zmiennej losowej, wyobraź sobie, że zdarzenie, które chcemy modelować, ma małe prawdopodobieństwo wystąpienia i że chcemy policzyć, ile razy zdarzenie wystąpiło w określonym przedziale czasowym. Jeśli mamy wyobrażenie o średniej liczbie wystąpień &mul, w określonym okresie czasu, podanej z poprzednich wystąpień, to zmienna losowa Poissona, oznaczona przez X = Poi(μ), zlicza całkowitą liczbę wystąpień wydarzenie w danym okresie. Innymi słowy, rozkład Poissona jest dyskretnym rozkładem prawdopodobieństwa, który zlicza liczbę zdarzeń występujących w danym przedziale czasu. Rozważ następujące przykłady zmiennych losowych Poissona:

• Znalezienie prawdopodobieństwa posiadania określonej liczby odwiedzających witrynę w ciągu godziny, znając wcześniejsze wyniki witryny

• Oszacowanie liczby wypadków samochodowych na skrzyżowaniu na podstawie wcześniejszych raportów policyjnych

Jeśli X = liczba zdarzeń w danym przedziale, a średnia liczba zdarzeń na przedział to λ liczby, to prawdopodobieństwo zaobserwowania x zdarzeń w danym przedziale określa wzór:

$$P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

Tutaj e = stała Eulera (2,718…).

Przykład - call center

Liczba połączeń przychodzących do call center jest zgodna z rozkładem Poissona w tempie 5 połączeń na godzinę. Jakie jest prawdopodobieństwo, że dokładnie sześć telefonów pojawi się między 22:00 a 23:00?

Aby skonfigurować ten przykład, wypiszmy podane przez nas informacje. Niech X będzie liczbą połączeń przychodzących między 22:00 a 23:00. To jest nasza zmienna losowa Poissona ze średnią λ = 5. Średnia wynosi 5, ponieważ używamy 5 jako naszej poprzedniej oczekiwanej wartości liczby połączeń przychodzących w tym czasie. Ta liczba mogła pochodzić z cennej pracy nad oszacowaniem liczby połączeń przychodzących co godzinę, a konkretnie po godzinie 22:00. Główną ideą jest to, że mamy pewne pojęcie o tym, ile połączeń powinno nadejść, a następnie wykorzystujemy te informacje do stworzenia naszej zmiennej losowej Poissona i używamy jej do przewidywania. Kontynuując nasz przykład, mamy:

P(X = 6) = = 0,146

Oznacza to, że istnieje około 14,6% szans na to, że dokładnie sześć połączeń nadejdzie między 22:00 a 23:00.

Skróty do oczekiwanej wartości Poissona i wariancji

Zmienne losowe Poissona mają również specjalne obliczenia dla dokładnych wartości oczekiwanych wartości i wariancji. Jeśli X jest zmienną losową Poissona ze średnią, to:

E(X) = λ

V(X) = λ

Jest to interesujące, ponieważ zarówno wartość oczekiwana, jak i wariancja to ta sama liczba, a ta liczba jest po prostu podanym parametrem! Teraz, gdy widzieliśmy już trzy przykłady dyskretnych zmiennych losowych, musimy przyjrzeć się drugiemu typowi zmiennej losowej, zwanej ciągłą zmienną losową.

Ciągłe zmienne losowe

Całkowite przełączanie biegów, w przeciwieństwie do dyskretnej zmiennej losowej, ciągła zmienna losowa może przyjmować nieskończoną liczbę możliwych wartości, a nie tylko kilka policzalnych. Nazywamy funkcje, które opisują krzywe gęstości rozkładu zamiast funkcji mas prawdopodobieństwa. Rozważ następujące przykłady zmiennych ciągłych:

• Długość rozmowy telefonicznej z przedstawicielem handlowym (nie liczba rozmów)

• Rzeczywista ilość oleju w beczce oznaczona 20 galonami (nie liczba beczek z olejem)

Jeśli X jest ciągłą zmienną losową, to istnieje funkcja f(x), taka, że dla dowolnych stałych a i b:

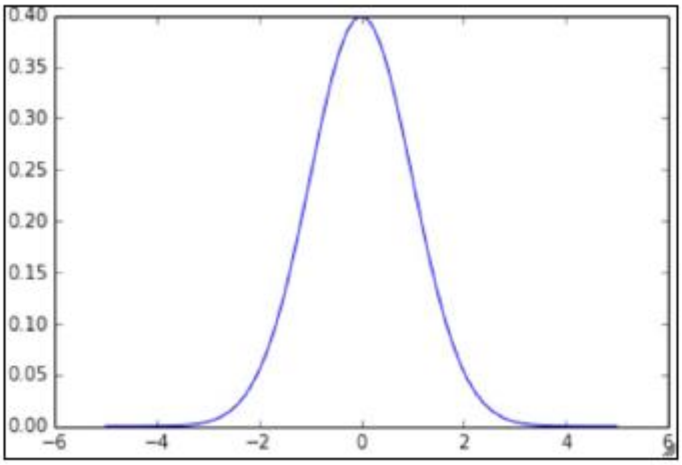
$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Poprzednia funkcja f(x) jest znana jako funkcja gęstości prawdopodobieństwa (PDF). PDF to wersja PMF z ciągłą zmienną losową dla dyskretnych zmiennych losowych. Najważniejszym rozkładem ciągłym jest standardowy rozkład normalny. Bez wątpienia albo słyszałeś o normalnym rozkładzie, albo miałeś z nim do czynienia. Idea stojąca za tym jest dość prosta. Plik PDF tej dystrybucji wygląda następująco:

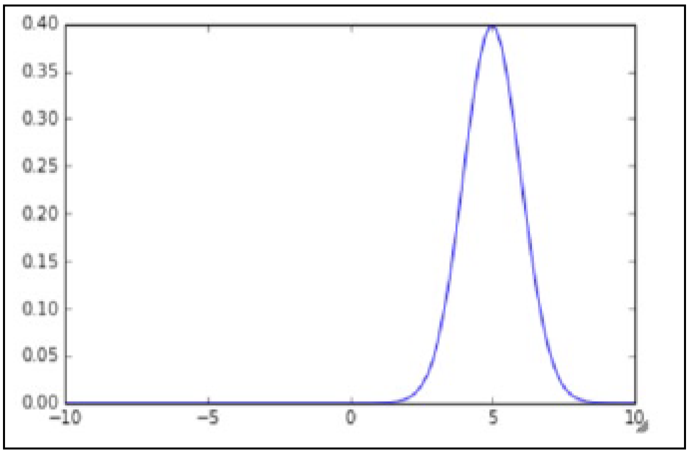
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Tu μ oznacza średnią zmiennej a σ odchylenie standardowe. Może to wyglądać na mylące, ale narysujmy to w Pythonie ze średnią równą 0 i odchyleniem standardowym równym 1, jak pokazano tutaj:

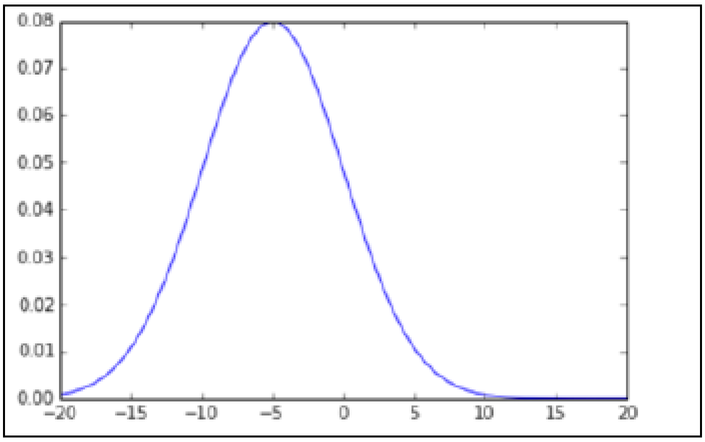
```
def normal_pdf(x, mu = 0, sigma = 1):

return (1./np.sqrt(2*3.14 * sigma**2)) * 2.718**(-(x-mu)**2 / (2.

* sigma**2))

x_values = np.linspace(-5,5,100)

y_values = [normal_pdf(x) for x in x_values]

plt.plot(x_values, y_values)
```



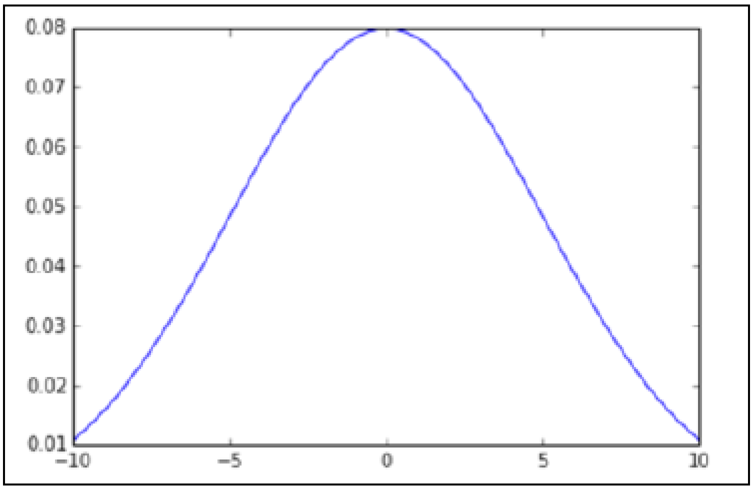
Co daje początek aż nazbyt znanej krzywej dzwonowej. Zauważ, że wykres jest symetryczny wokół linii x = 0. Spróbujmy zmienić niektóre parametry. Najpierw spróbujmy z μ = 5 :



Następnie spróbujmy z wartością σ = 5 :



Na koniec spróbujemy z wartościami μ = 5 i σ = 5 :



Na wszystkich wykresach mamy standardowy kształt dzwonka, który wszyscy znamy, ale gdy zmieniamy nasze parametry, widzimy, że dzwonek może stać się cieńszy, grubszy lub przesunąć się od lewej do prawej. W następnych rozdziałach, które skupiają się na statystyce, w dużo większym stopniu wykorzystamy rozkład normalny, ponieważ odnosi się on do myślenia statystycznego.

Podsumowanie

Prawdopodobieństwo jako pole wyjaśnia nasz losowy i chaotyczny świat. Korzystając z podstawowych praw prawdopodobieństwa, możemy modelować rzeczywiste zdarzenia z udziałem losowości. Możemy użyć zmiennych losowych do reprezentowania wartości, które mogą przyjmować kilka wartości, a funkcji masy lub gęstości prawdopodobieństwa możemy użyć do porównania linii produktów lub przyjrzenia się wynikom testów. Widzieliśmy niektóre z bardziej skomplikowanych zastosowań prawdopodobieństwa w przewidywaniu. Korzystanie ze zmiennych losowych i twierdzenia Bayesa to doskonałe sposoby przypisywania prawdopodobieństw do sytuacji z życia codziennego. W kolejnych rozdziałach ponownie przyjrzymy się twierdzeniu Bayesa i wykorzystamy je do stworzenia bardzo potężnego i szybkiego algorytmu uczenia maszynowego, zwanego algorytmem Naïve Bayes. Algorytm ten wychwytuje moc myślenia bayesowskiego i stosuje go bezpośrednio do problemu uczenia się predykcyjnego. Kolejne dwa rozdziały koncentrują się na myśleniu statystycznym. Podobnie jak prawdopodobieństwo, rozdziały te będą wykorzystywać formuły matematyczne do modelowania rzeczywistych wydarzeń. Główną różnicą będzie jednak terminologia, której używamy do opisywania świata oraz sposób, w jaki modelujemy różnego rodzaju zdarzenia. W kolejnych częściach postaramy się zamodelować całe populacje punktów danych wyłącznie na podstawie próbki. Z prawdopodobieństwem ponownie przyjrzymy się wielu pojęciom, aby nadać sens twierdzeniom statystycznym, ponieważ są one ściśle powiązane i oba są ważnymi pojęciami matematycznymi w dziedzinie nauki o danych