Module 2 - Generative Modeling I

2.1 The Generative Approach to Classification

2.2 Probability Review I: Probability Spaces, Events, and Conditioning

2.3 Generative Modeling in One Dimension

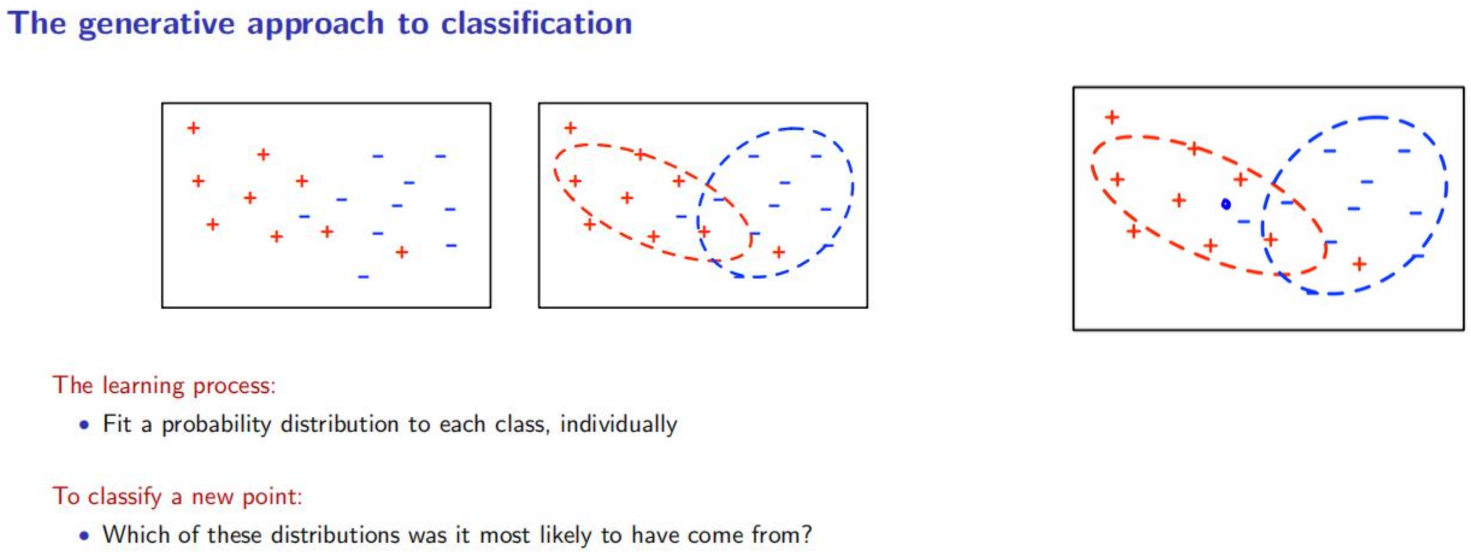
2.4 Probability Review II: Random Variables, Expectation, and Variance

2.5 Probability Review III: Measuring Dependence

2.6 Two Dimensional Generative Modeling with the Bivariate Gaussian

总结：

2.1 The Generative Approach to Classification



The generative approach to classification

The learning process:

- Fit a probability distribution to each class, individually

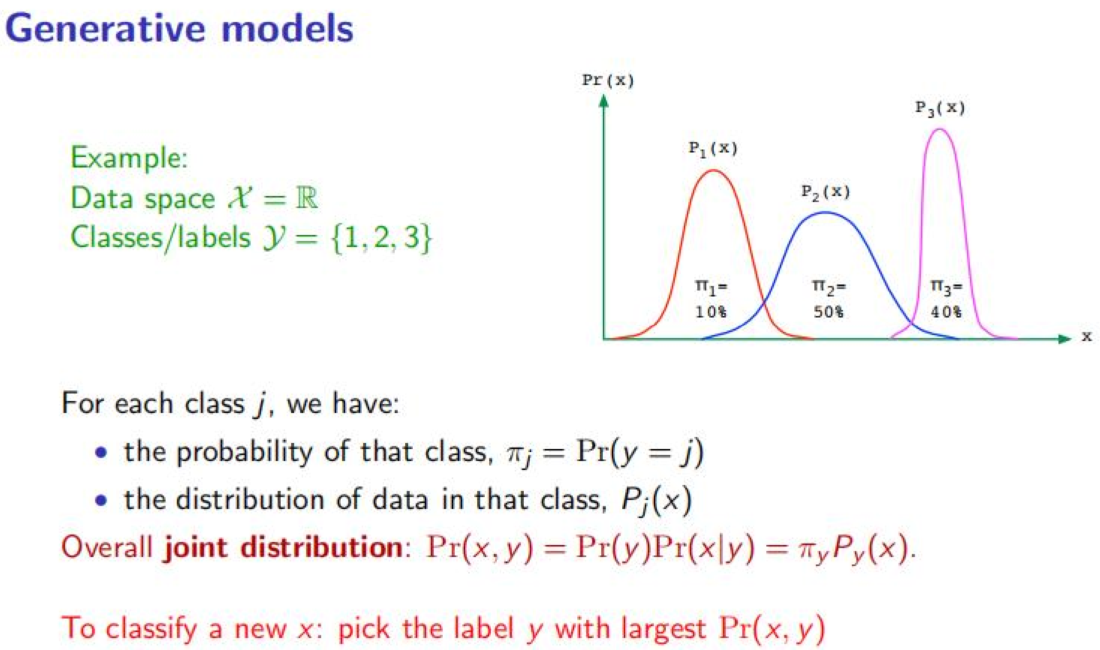
To classify a new point:

- Which of these distributions was it most likely to have come from?

Today we will introduce a simple and powerful approach to building classifiers. It's called the generative approach and it's based on probability distributions. The main idea with the generative approach is to fit each class separately with a probability distribution.

So for example, here we have a training set of about 15-20 points and what we do is that we first look only at one label, so here there are two labels, pluses and minuses. So we start by looking just at the pluses and we fit a model to them. And then we look only at the minuses, and we fit a model to those. So maybe we get something like this. So on the left, we have an ellipse shape distribution that we fit to the pluses and then we have another ellipse shaped distribution that we fit into the minuses. That's a total learning process.

Now, when a new point comes along, say, a point like this, and we want to classify it, we just ask ourselves Was this new point more likely to have come from the red distribution or the blue distribution? Under which of these two distributions does it have higher probability? And that's it. Okay, so that's a high level overview of the generative approach.



Generative models

Example:
Data space $\mathcal{X} = \mathbb{R}$
Classes/labels $\mathcal{Y} = \{1, 2, 3\}$

For each class j, we have:

- the probability of that class, $\pi_j = \Pr(y = j)$
- the distribution of data in that class, $P_j(x)$

Overall **joint distribution**: $\Pr(x, y) = \Pr(y)\Pr(x|y) = \pi_y P_y(x)$.

To classify a new x: pick the label y with largest $\Pr(x, y)$

So let's be a little bit more concrete and put down some details. Here's a picture in which there are three classes, or three labels. Let's call them one, two, and three. So the label space y is just the set one, two, and three. And let's say that the data is very simple. The data just has one feature, so the data points are one-dimensional and they can just be plotted on the line.

- So in order to fit a generative model, what we do, is we take our training set and first we look only at the points whose label is one, and we fit a probability distribution to them. So maybe, for example, we get this distribution over here, the red one. Let's call that one P1(x). The distribution for label one. (续下一页)

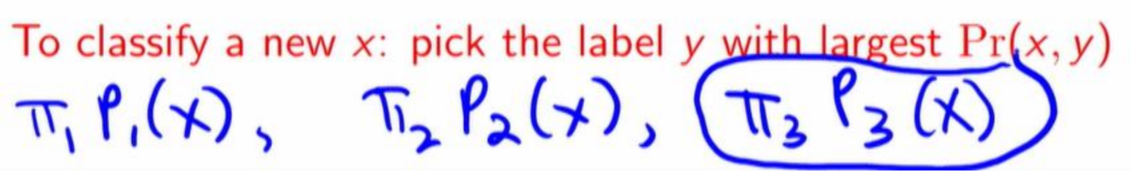
总结：

- Then, we take the training set again, and look only at the points whose label is two, and we fit a probability distribution to those. Maybe that looks like P2(x) over here.
- And finally we look at the remaining points, the ones that have labeled three, and we get a probability distribution to that. So that's P3(x).

And now we have these three separate probability distributions. We need a little bit more information, as well.

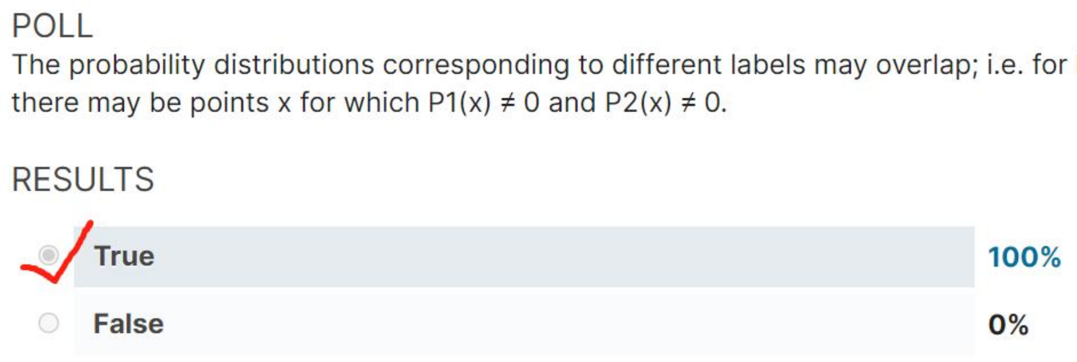
- Let's say, for instance, that label one makes up 10% of the training set. So out of the training points, one tenth of them have label one. So what we'll do is say Pi(1) equals 10% or .1. Let's say label two makes up 50% of the training set. So we'll say Pi(2) equals 50% or .5 and the remaining 40% of the training set comes from label three. So, we say Pi(3) equals .4.
- Now, at the end of this, what we have is the following pieces of information. For each class, j, so j is one, two, or three, we have the probability of that class, which is just a number, like .1, .5, or .4, and we also have the distribution of data in that class, which is what we have been calling P sub j, P set 1, P set 2, P set 3. And these pieces are built information are enough to fully specify the joint distribution between x and y.
- So the probability of any pair, x y, is the probability of that label, y, times the probability of seeing data x under that label y. So the probability of x given y. Now the probability of seeing label y is simply Pi sub y and the probability of seeing point x under the distribution of label y, is P sub y of x. So we have the full probability distribution over x and y. So now a new point x comes along and we want to classify it. We want to determine a label, y, for it. Which label do we pick? Well, **the mostly likely label, the one that maximizes the probability of x and y**.
    - Concretely, what we would do in this case, with three classes, is that we take our point x and we calculate Pi one times P1(x) and that's some number. We calculate Pi two times P2(x) and we calculate Pi three times P3(x). And we take whichever of these is the largest. If the last one is the largest, for example, we say the label equals three.

So that's it for a high-level overview of the generative approach to classification. Some of the details might seem a little mysterious at this point because they rest on various probabilistic concepts.
So what we'll do next is to do a little bit of a review, a little bit of a tour, of the relevant probability and then we'll come back and we'll flesh out this approach and see it in action.



To classify a new x: pick the label y with largest $\Pr(x, y)$

$\pi_1 P_1(x)$, $\pi_2 P_2(x)$, $\boxed{\pi_3 P_3(x)}$



POLL

The probability distributions corresponding to different labels may overlap; i.e. for labels 1 and 2, there may be points x for which P1(x) ≠ 0 and P2(x) ≠ 0.

RESULTS

● True ✓ 100%

○ False 0%

总结：

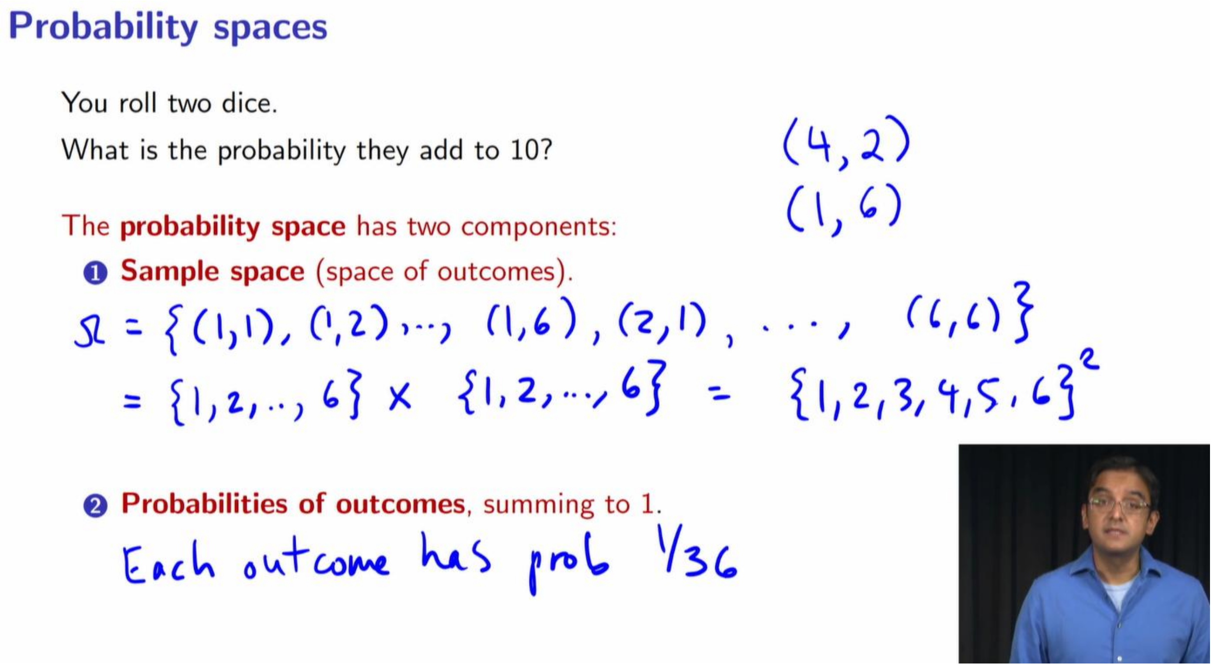
2.2 Probability Review I: Probability Spaces, Events, and Conditioning

Topics we'll cover

1. How to define the **probability space** for an experiment in which outcomes are random.
2. How to formulate an **event** of interest.
3. The probability that two events both occur.
4. The **conditional probability** that an event occurs, given that some other event has occurred.
5. **Bayes' rule**.

In order to progress with machine learning there are three types of math that you need to get pretty comfortable with, probability, linear algebra, and optimization. Of these three, I'm guessing that probability is the one that most of you are probably the most familiar with. But just to be sure, what I'd like to do is a little bit of a review broken down into three parts, and by the time we're done with this we'll be ready to delve much more deeply into generative modeling for classification.

So what we'll do today is to define the notion of the probability space for an experiment. We'll see how to formulate an event that interests us, and we'll see how one event influences the probability of another event. All of this will lead ultimately to Bayes' rule, which is the central formula for probabilistic reasoning in machine learning and statistics.



Probability spaces

You roll two dice.

What is the probability they add to 10?

$(4,2)$

$(1,6)$

The **probability space** has two components:

1. **Sample space** (space of outcomes).

$$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$$
$$= \{1,2,..,6\} \times \{1,2,\ldots,6\} = \{1,2,3,4,5,6\}^2$$

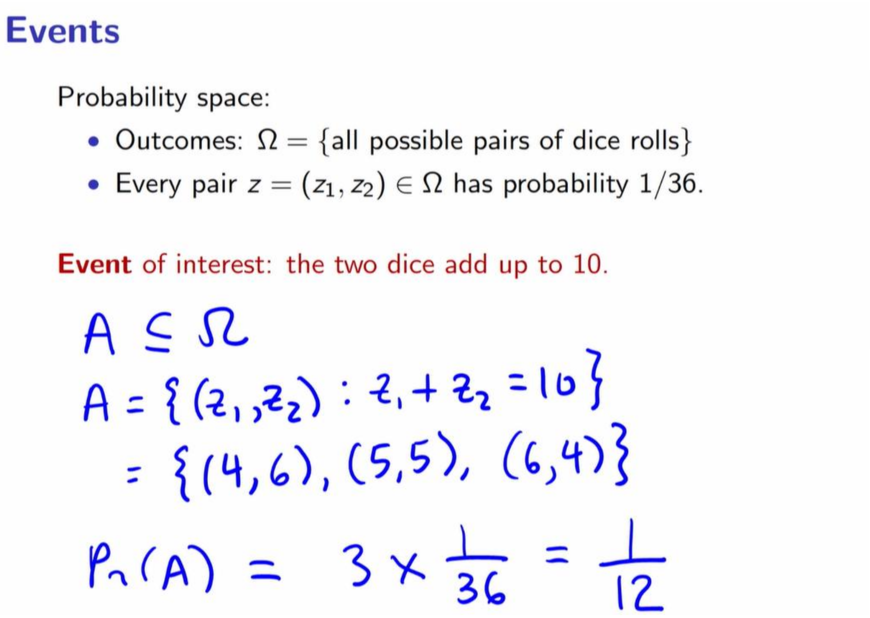
2. **Probabilities of outcomes**, summing to 1.

Each outcome has prob 1/36

So let's start with **probability spaces**. A probability space for a random experiment is a summary of all the information we need in order to answer questions about the experiment. Let's hear a concrete example. So suppose we rolled two dice and we wonder, what is the probability that they're gonna add up to 10? So it's a random experiment, and the probability space for this experiment has two components to it, the set of all possible outcomes, which is called the **sample space**, and the probabilities of each of these outcomes.

- So what are the outcomes for this experiment? Well we rolled two dice so maybe the first one is a four and the second one is a two. We'll write it like that. Or maybe the first one is a one and the second die turns out to be a six. The sample space is the set of all possible outcomes, so it's the set of all possible pairs like this. So the sample space consists of one, one, one, two, all the way to one, six, then two, one, and so on and so forth all the way to two sixes. These are all the possible outcomes, the set of all possible outcomes. We could also write this a little bit more consistent, a little bit more concisely as all possible outcomes for the first die, set product with all possible outcomes for the second die. And in fact an even more concise way of writing it is just as one, two, three, four, five, six, squared. Okay, so that's the sample space.
- What are the probabilities of these various outcomes? So the number of outcomes, the number of possible outcomes is six times six, 36, and they are all equally likely. And so each outcome has probability 1/36.

总结：



Events

Probability space:

- Outcomes: $\Omega = \{\text{all possible pairs of dice rolls}\}$
- Every pair $z = (z_1, z_2) \in \Omega$ has probability $1/36$.

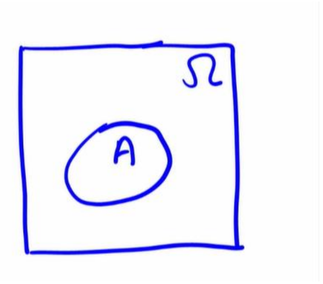
Event of interest: the two dice add up to 10.

$$A \subseteq \Omega$$

$$A = \{(z_1, z_2) : z_1 + z_2 = 10\}$$

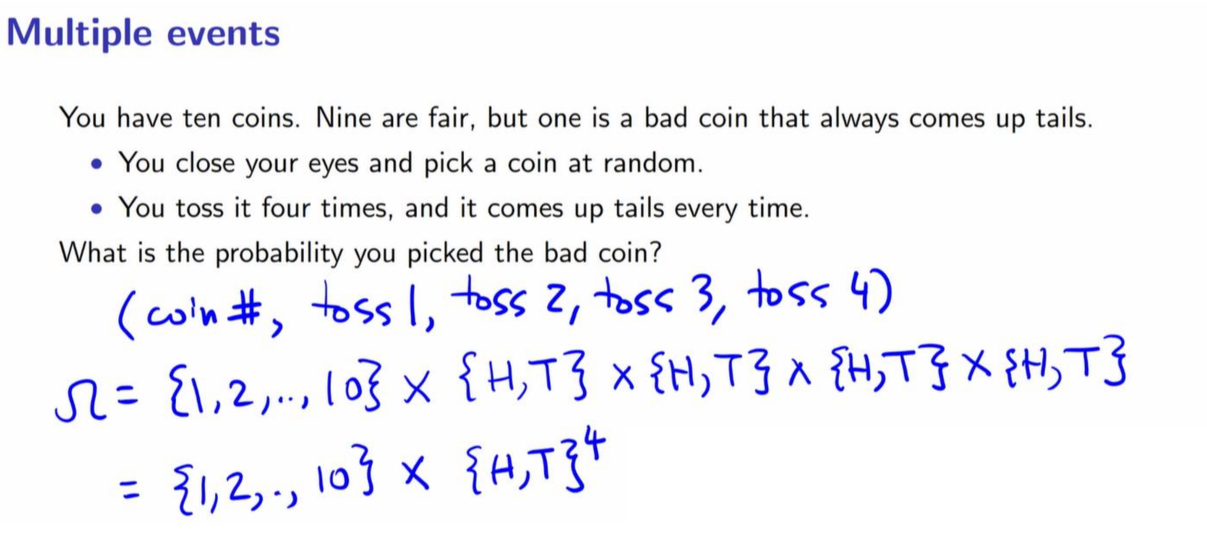
$$= \{(4,6), (5,5), (6,4)\}$$

$$Pr(A) = 3 \times \frac{1}{36} = \frac{1}{12}$$



So now that we have our probability space, let's formulate the event of interest. And so the event we care about is when the two dice add up to 10, and we're wondering what is the probability of this event? So let's call the event A. An event is just a subset of the possible outcomes. So it's a subset of the possible outcomes. And in particular, the event consists of all pairs, z1, z2 let's say, that add up to 10. Okay, so which pairs are these?

Well if the first die is a one then the second one has to be a nine, so that doesn't work. So the first die actually has to be at least a four. So if the first one is a four then the second one has to be a six, if the first one is a five then the second one has to be a five, and if the first one is a six then the second one has to be a four, and that's it. So this is the event that interests us. It consists of 3 possible outcomes out of the 36. If we were to draw a Venn diagram for this we have the set of all 36 outcomes and the event that interests us, which consists of just three out of these events. And what is the probability of it? Well the three outcomes, and each of them has probability 1/36, and so the probability of A is just 1/12. So that's how we define an event and compute its probability.



Multiple events

You have ten coins. Nine are fair, but one is a bad coin that always comes up tails.

- You close your eyes and pick a coin at random.
- You toss it four times, and it comes up tails every time.

What is the probability you picked the bad coin?

(coin #, toss 1, toss 2, toss 3, toss 4)

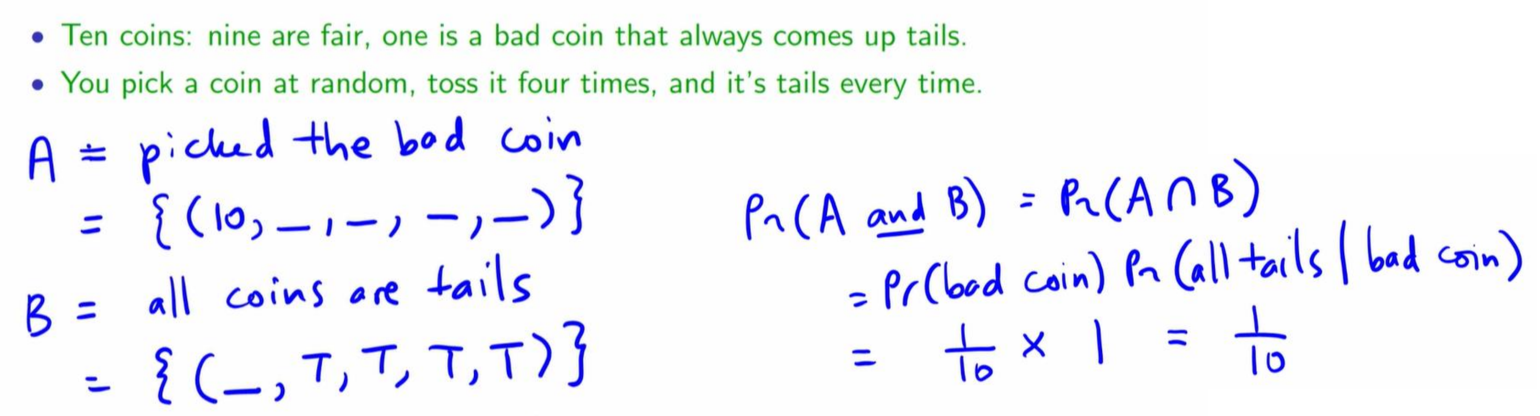
$$\Omega = \{1, 2, .., 10\} \times \{H, T\} \times \{H, T\} \times \{H, T\} \times \{H, T\}$$

$$= \{1, 2, ., 10\} \times \{H, T\}^4$$

Now when we are using probabilities in statistics and machine learning, we're typically dealing not just with one event but with **multiple events**. For instance, we measure a patient's temperature and blood pressure, and wonder whether the patient has a particular disease. So that's a coming together of three events. There's the event that the patient has that particular blood pressure, there's the event that the patient has a particular temperature, and there's the event that the patient has the disease. So we get to observe two of these events, the temperature and the blood pressure, and we wonder what they tell us probabilistically about the third event, the one we don't get to observe, the one we care about, the disease.

- Let's see a concrete example of this type. Here's a toy situation. You have 10 coins in front of you and they all look the same but they're not the same. 9 of them are regular fair coins, coins that come up heads half the time and tails half the time, but the tenth coin is a bad coin that always comes up tails. So you close your eyes and you pick a coin at random, and then you toss the coin four times and you find that it comes up tails every single time. What is the probability that you picked the bad coin? So this is something that's not obvious and we're gonna have to work this out.
- First of all, what is the sample space for this random experiment? So what's going on? We first pick a coin at random and then we toss it four times. Okay, so the way I'm going to write down an outcome is first I'll say which coin we picked, just the coin number, coin number one to 10. Let's say 10 is the bad coin. And then what happened on the first toss, heads or tails, and the second toss, the third toss, and the fourth toss. That's how we'll describe a particular outcome, and so the set of all possible outcomes is the set of all choices for the coin, there are 10 possible choices, product with the set of all possibilities for toss one and toss two, and three and four. This is the sample space of possible outcomes. In fact, we can write this a little bit more concisely. So we can also write it as one, all coin choices, times H T to the fourth. So how many possible outcomes are there? Well there's 10 times two times two times two times two, so 160 possible outcomes. Okay, so that's the sample space.

总结：



- Ten coins: nine are fair, one is a bad coin that always comes up tails.
- You pick a coin at random, toss it four times, and it's tails every time.

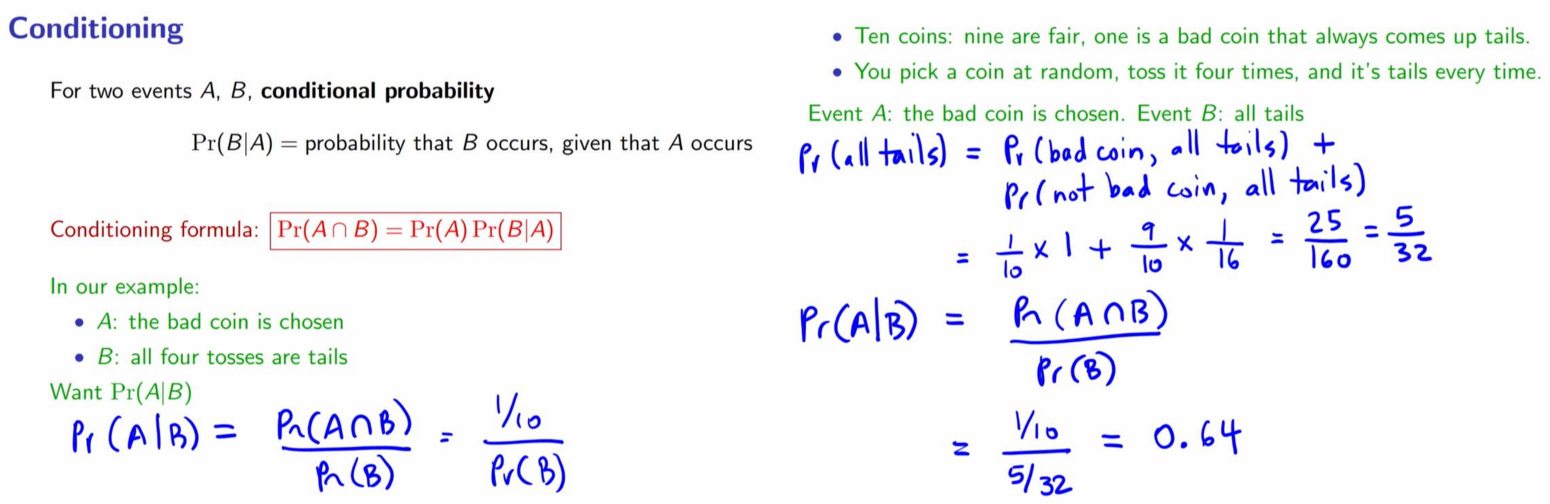
$$A \doteq \text{picked the bad coin} = \{(10, \_, \_, \_, \_)\}$$

$$B = \text{all coins are tails} = \{(\_, T, T, T, T)\}$$

$$\Pr(A \underline{\text{ and }} B) = \Pr(A \cap B) = \Pr(\text{bad coin}) \Pr(\text{all tails} \mid \text{bad coin}) = \frac{1}{10} \times 1 = \frac{1}{10}$$

- So now the event we care aboutis whether or not we got the bad coin. So let's call that A. So A is the event that we picked the bad coin. And formally what the event is, is the set of all outcomes in which the coin we picked was the bad one, coin number 10, and then it doesn't matter what happens afterwards. So this is the event we care about, but it's not all we get to observe.
- We get to observe a different event which is related, and that's that all the coins come up tails. And so here we don't know what the coin number is, but all the outcomes are tails.

So these are the two events that we're dealing with. What is the probability that both of them occur? So what is the probability, A and B? And the way we usually write this, in fact, is A intersection B. What is the probability that both of these things happen? Well it's the probability that we picked the bad coin times the probability that we get all tails given the bad coin, given that the coin is bad. The probability of picking the bad coin is 1/10 since there's one bad coin and nine good coins. And once we've picked the bad coin the probability that all four tosses are gonna come up tails, that is one. So the probability that both A and B happen is 1/10.



Conditioning

For two events A, B, **conditional probability**

$$\Pr(B|A) = \text{probability that } B \text{ occurs, given that } A \text{ occurs}$$

Conditioning formula: $\boxed{\Pr(A \cap B) = \Pr(A)\Pr(B|A)}$

In our example:

- A: the bad coin is chosen
- B: all four tosses are tails

Want $\Pr(A|B)$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/10}{\Pr(B)}$$

- Ten coins: nine are fair, one is a bad coin that always comes up tails.
- You pick a coin at random, toss it four times, and it's tails every time.

Event A: the bad coin is chosen. Event B: all tails

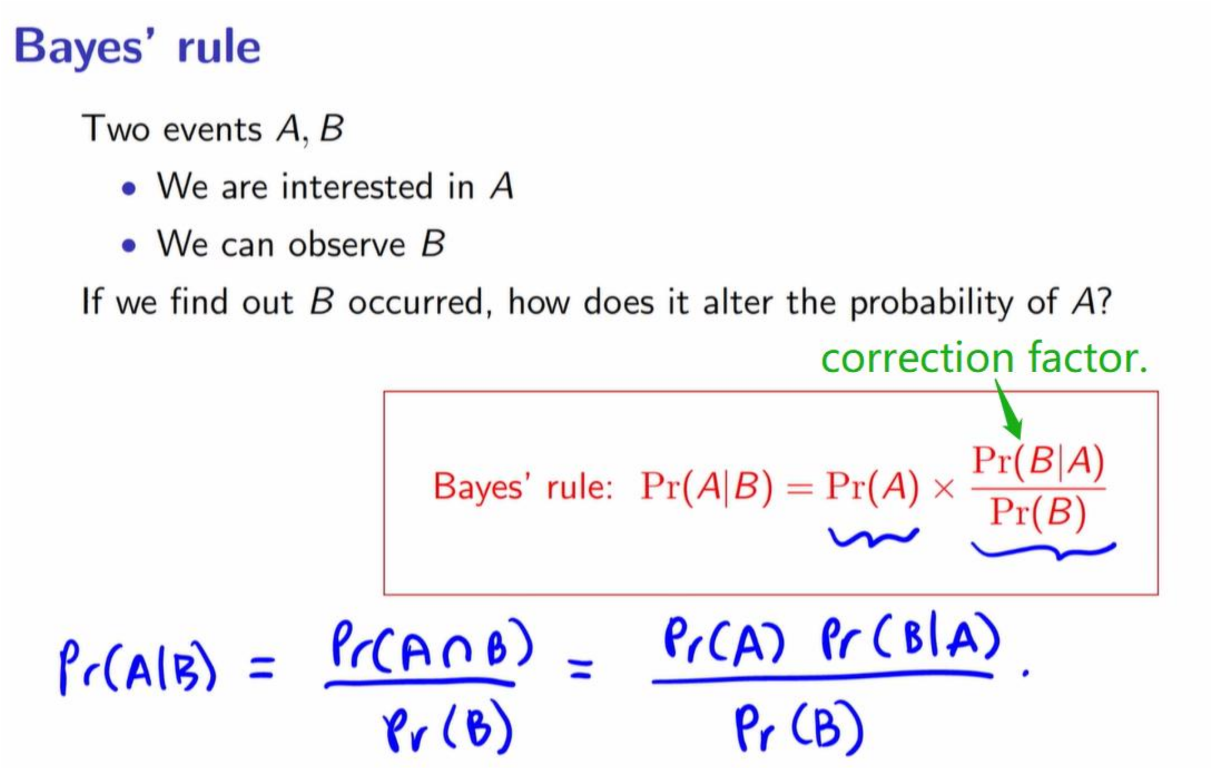
$$\Pr(\text{all tails}) = \Pr(\text{bad coin, all tails}) + \Pr(\text{not bad coin, all tails}) = \frac{1}{10} \times 1 + \frac{9}{10} \times \frac{1}{16} = \frac{25}{160} = \frac{5}{32}$$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/10}{5/32} = 0.64$$

So this has all been fairly intuitive but we are using some general rules about conditioning over here, and so let's be a little bit more explicit about that. So when we have two events, A and B, we can talk about the conditional probability of B given A, the probability that B occurs given that we know A occurs. And the notation for that involves this bar over here. This bar means given. Now the most basic formula for conditional probability is this one shown over here. It says that the probability that A and B both occur is the probability that A occurs times the probability that B occurs given that A has occurred.

- So returning to our example, there are two events that we're dealing with, A and B. A is that the bad coin is chosen, B is that we have all tails, and we are interested in the probability of A given B. We've seen that B is true and we're wondering what is the probability of A given this information. Well we can compute it directly from this formula. So if we rearrange terms we see that the probability of A given B is the probability of A and B divided by the probability of B.
- So what is the probability of A and B? We've computed that, that's 1/10.
- What is the probability of B? What is the probability that all four tosses come up tails? Well that's something we have to figure out. Okay, so let's go ahead and do that. So the probability that all four tosses are tails, let's figure that out. Well it kind of depends on whether or not we chose the bad coin, so let's break it up into those two cases. So it's the probability that we got the bad coin and we get all tails, plus the probability that we did not get the bad coin and yet we still get all tails. So now let's compute these. The probability that we get the bad coin is 1/10. If we get the bad coin, the probability of all tails is one because the bad coin always comes up tails. The probability that we don't get the bad coin is 9/10 because there are nine good coins. And if we don't get the bad coin, if we have a fair coin, the probabilityof getting all tails, of getting four tails in a row, is 1/2 times 1/2 times 1/2 times 1/2, so it's 1/16. Okay, so let's work this out. 1/10 plus 9/160 is 25/160, which is 5/32. So this is the probability of B, the probability of all tails, and now let's go back to what we actually wanted, the probability of A given B.
- And as we saw, it's the probability of A and B over the probability of B. The probability of A and B we figured out, that was 1/10. The probability of B we just figured out, it's 5/32. Okay, and so this works out to 32/50, which is 64/100 or 0.64. So there is exactly a 64% probability that we have the bad coin. Pretty impressive, huh?

总结：



Bayes' rule

Two events A, B

- We are interested in A
- We can observe B

If we find out B occurred, how does it alter the probability of A?

correction factor.

$$\boxed{\text{Bayes' rule:} \quad \Pr(A|B) = \Pr(A) \times \frac{\Pr(B|A)}{\Pr(B)}}$$

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)\ \Pr(B|A)}{\Pr(B)}.$$

So when we were doing those calculationswe were implicitly using **Bayes' rule**, which is the central formula for probabilistic reasoning in statistics and machine learning.

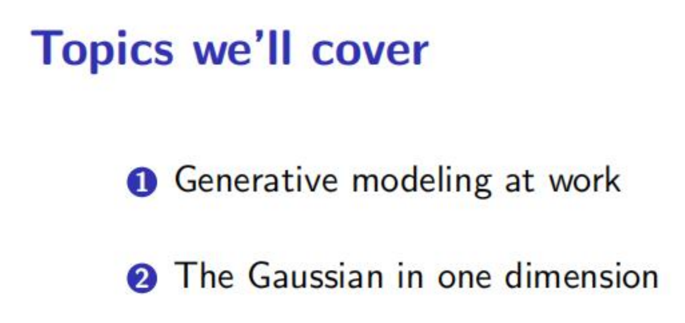
So let's spell out this formula a little bit more precisely now. We have two events, A and B, and the event we are interested in is event A, but we don't get to observe it directly. We get to observe some other event B, and what we're interested in is given that B happened, what does it tell us about the probability of A happening? So here is Bayes' rule. **If we don't observer anything about B, the probability that A happens is just probability of A**, but now that we know that B happened, that changes the probability of A. By how much? By a multiplicative **correction factor**, and this is the correction factor. Now this formula is just two applications of the general rule for conditioning. So let's see how we did it.

- The probability of A given B is as we know, just the probability of A and B over the probability of B. And now let's just work on the numerator. The probability of A and B is just the probability of A times the probability of B given A. And we keep the denominator the same, and sure enough we have Bayes' rule.

The various calculations we've done today might seem very simple but they're actually very powerful. Bayes' rule tells us how we can reason under uncertainty, and it's no surprise, therefore, that it's one of the fundamental formulas for inference in machine learning.

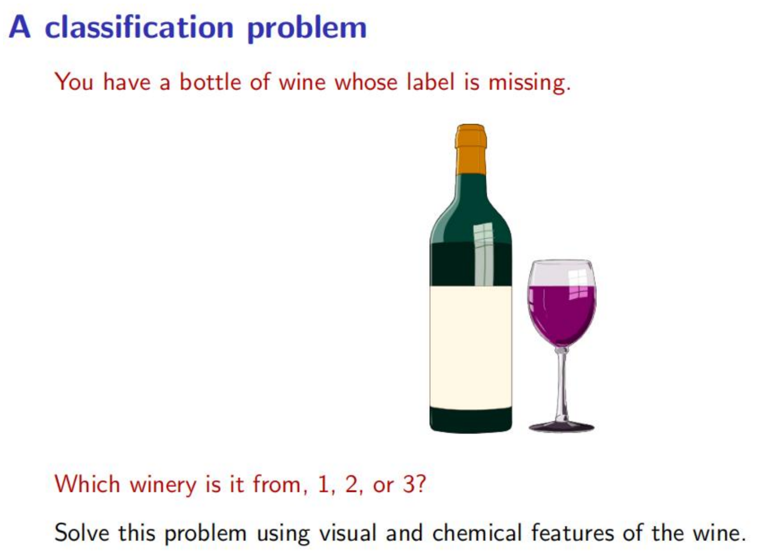
总结：

2.3 Generative Modeling in One Dimension



Topics we'll cover

1. Generative modeling at work
2. The Gaussian in one dimension



A classification problem

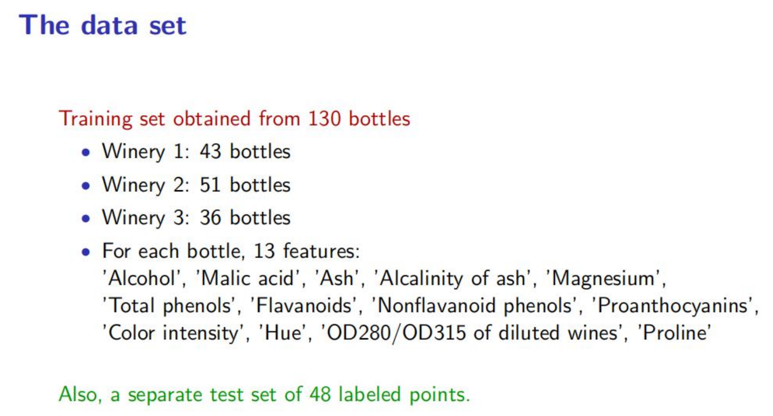
You have a bottle of wine whose label is missing.

Which winery is it from, 1, 2, or 3?

Solve this problem using visual and chemical features of the wine.

So we've talked a little bit about the generative approach to classification and today, we'll put it to use. We'll look at a simple one-dimensional data set, which has three classes in it. And we'll see how we can classify this data, by fitting a Gaussian distribution to each of these classes.

So, here's the problem that we are dealing with. We have a bottle of wine, and we like it, but its label is missing. And so we want to figure out which winery it is from. One, two or three? We're gonna take the machine learning approach and we'll measure some visual and chemical features of the bottle of wine and use this to predict the label, one, two or three.

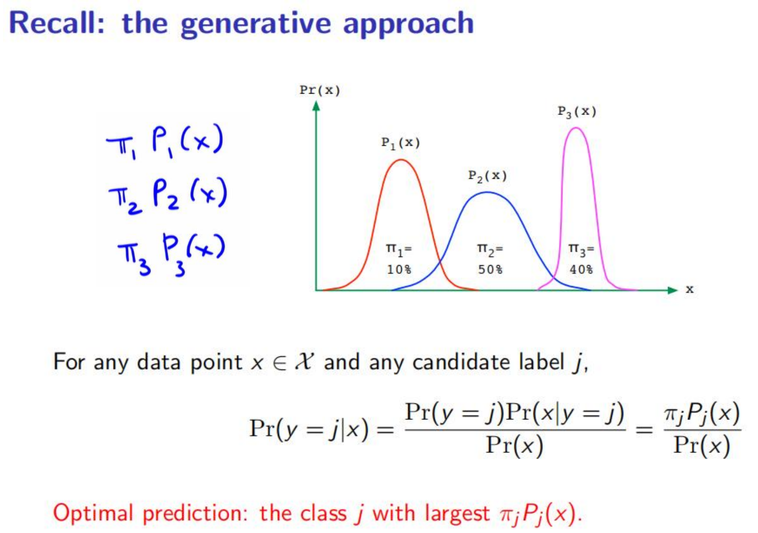


The data set

Training set obtained from 130 bottles

- Winery 1: 43 bottles
- Winery 2: 51 bottles
- Winery 3: 36 bottles
- For each bottle, 13 features:
  'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash', 'Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'

Also, a separate test set of 48 labeled points.



Recall: the generative approach

For any data point $x \in \mathcal{X}$ and any candidate label j,

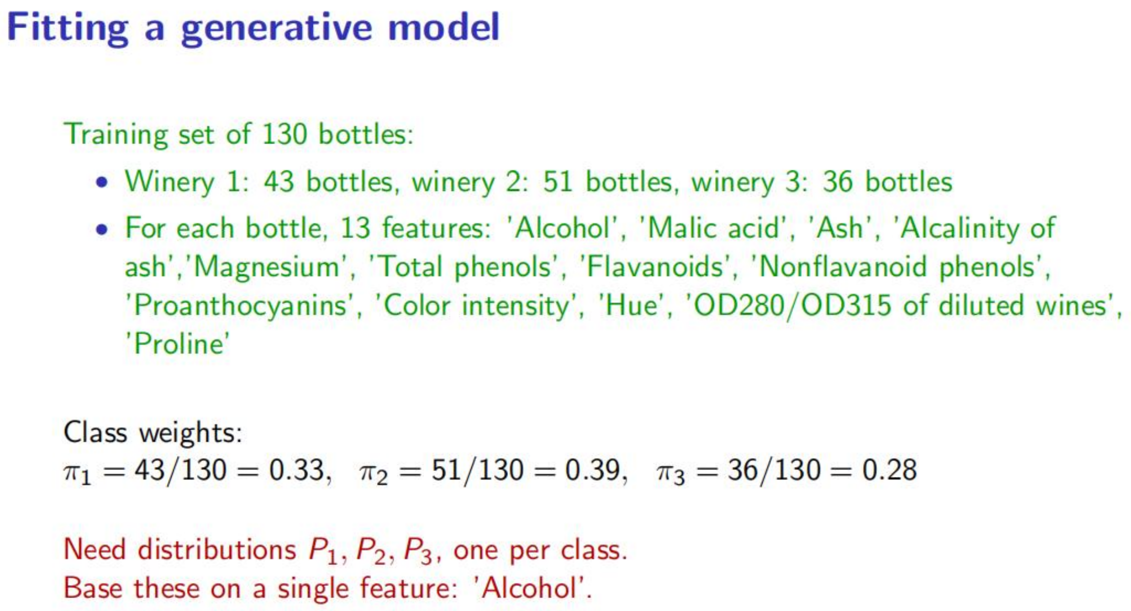
$$\Pr(y=j|x) = \frac{\Pr(y=j)\Pr(x|y=j)}{\Pr(x)} = \frac{\pi_j P_j(x)}{\Pr(x)}$$

Optimal prediction: the class j with largest $\pi_j P_j(x)$.

Now, we have our training set to help us and this training set was obtained from 130 bottles, 43 of these bottles were from winery one, 51 from winery two, and the remaining 36 were from winery three. For each of these bottles, a data point was measured consisting of 13 features. Now, these are features that makes sense to wine experts, things like Flavanoids and Proline and Magnesium levels. But there are 13 said features and so the data points are essentially 13 dimensional vectors. So, we're gonna use this data to build a classifier that takes the features from a new bottle and predicts the label one, two or three. And there's also a separate test set with 48 labeled points that we can use to evaluate how good our classifier is.

Let's quickly recall the generative approach,the classification. So, what we need to do here is to fit a distribution to each winery individually. So, for winery one, we fit some distribution, say, p one of x. For winery two, we fit a distribution, p two. And we have a distribution, p three, which captures the data from winery three. We also need to know the probabilities of each of these labels. So, what percentage of the data, what percentage of bottles come from winery one, that's π1. What percentage of bottles come from winery two, that's π2 and π3. So these π's add up to one. And once we have all this information, that's our model, when we get a new bottle of wine, x, we get its feature vector, that's the x, and the class we predict one, two or three is simply the class that maximizes πj times pj of x. So, in this case, because there are three classes, we simply compute π1 times p one of x, and then we compute π2 times p two of x, and π3 times p three of x. And we take whichever of these is the largest.
Why do we do that exactly? Well, this is a direct consequence of Bayes' rule. So, we have this new bottle, whose features are x. We want to know what its label is. And so, the label we are gonna pick is the most likely label given x. So, what is the probability that the label is j, given the specific features x. Using Bayes' rule, it's the probability of having label j times the probability of seeing x in label j divided by the overall probability of seeing x. So, the probability that the label of j is just π sub j. The probability that we would see x amongst the bottles from winery j is pj of x **and the denominator doesn't depend on j at all, so we can ignore it**. So, we are simply gonna pick the label j that maximizes this equation over here. That's the generative approach.

总结：



Fitting a generative model

Training set of 130 bottles:

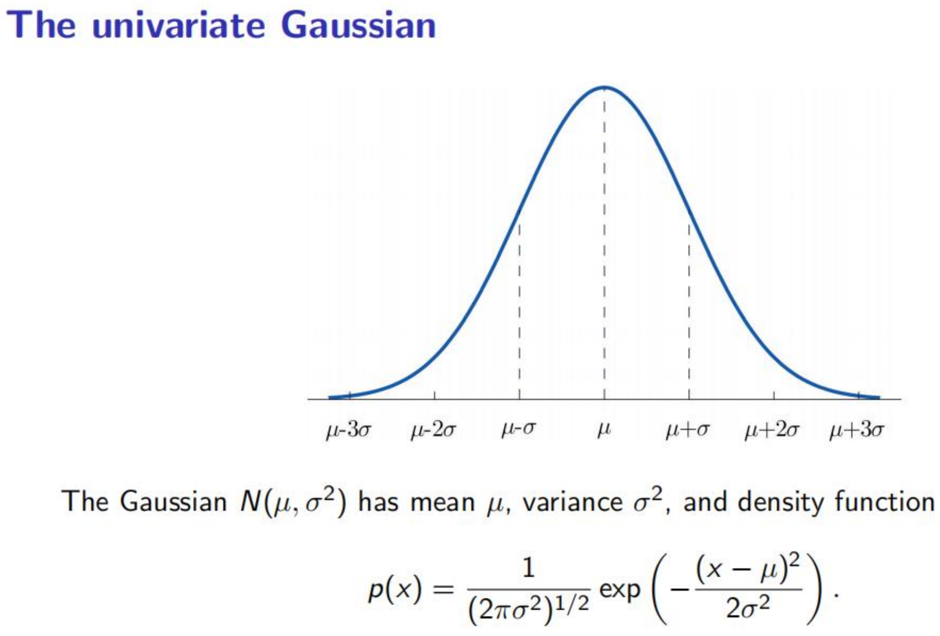
- Winery 1: 43 bottles, winery 2: 51 bottles, winery 3: 36 bottles
- For each bottle, 13 features: 'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash','Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'

Class weights:
$\pi_1 = 43/130 = 0.33, \quad \pi_2 = 51/130 = 0.39, \quad \pi_3 = 36/130 = 0.28$

Need distributions P_1, P_2, P_3, one per class.
Base these on a single feature: 'Alcohol'.

So, let's go back to the data now. This is the summary of our data set, and the first thing we need to do is to figure out the probabilities of each of the individual classes. So, out of the 130 bottles, 43 of them are from winery one. That means the probability of winery one, the weight of winery one is 43/130. So, π sub one, the weight of winery one is 43/130, which is about 0.33. Out of the 130 training bottles, 51 are from winery two. So, the probability of winery two is 51/130, which is 0.39. And likewise, the probability of winery three is 36/130, which is 0.28. And as you can see that these three numbers add up to one.

So, that part's easy and in general, fitting the class weights is very easy, almost trivial. The harder part is fitting a distribution to each class. Now, in this case, the data consist of 13 features. So, it's a 13-dimensional data set. So, we need a distribution over 13 dimensional space. We haven't yet seen how to do that. And so, let's just simplify things for the time being and just pick one feature out of the 13. So, we have these 13 features, we're just gonna pick one of them. Let's just go ahead and choose alcohol level as our feature. And so we've now reduced the data to just one dimension and we're curious, how well can we predict the winery using alcohol level alone. Good, so now we need to fit one distribution to the alcohol levels from winery one. Another distribution to the alcohol levels from winery two and a third distribution to the alcohol levels from winery three.



The univariate Gaussian

The Gaussian $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

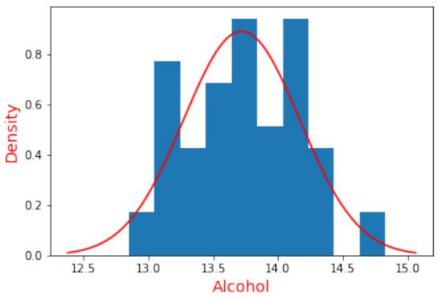
$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

What one dimensional distribution should we use? Well, the default choice is the Gaussian. So, let's look at it. The Gaussian in one dimension is just the familiar bell curve. It's a distribution that is specified by just two parameters, a mean and a variance. I'm gonna be denoting the mean by the Greek letter, mu. And the variance by the Greek letter, sigma squared. So the mean is mu, I'm just gonna emphasize it by rewriting it. And the variance is sigma squared, which means that the standard deviation is sigma. We're gonna refer to this distribution by using the short hand n of mu, sigma squared. The Gaussian with mean, mu and variance sigma squared. The N stands for normal and people call it that because a lot of data looks like this. Now, this particular distribution is a distribution of real numbers. Its density is given over here by this formula. The density might look a little bit messy or complicated at this point, but believe me it will start seeming a lot more familiar as time goes on.
So, what can we say about the distribution? So, first of all, it is centered at the mean mu, that's right here. It's symmetric about the mean. About 2/3 of the distribution lies within one standard deviation of the mean. About 95% of the distribution lies within two standard deviations of the mean. So, it's a distribution that's fairly well concentrated about its mean, and these are the distributions that we're gonna be using.

总结：

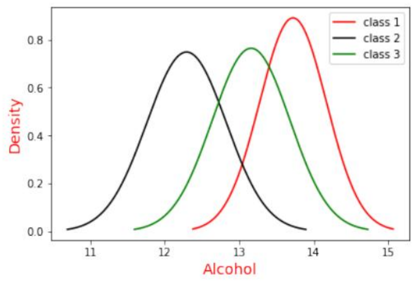
The distribution for winery 1

Single feature: 'Alcohol'



Mean $\mu = 13.72$, Standard deviation $\sigma = 0.44$ (variance 0.20)

All three wineries



- $\pi_1 = 0.33,\ P_1 = N(13.7, 0.20)$
- $\pi_2 = 0.39,\ P_2 = N(12.3, 0.28)$
- $\pi_3 = 0.28,\ P_3 = N(13.2, 0.27)$

To classify x: Pick the j with highest $\pi_j P_j(x)$

Test error: 14/48 = 29%

So, we need to fit a Gaussian to the datafrom each of the wineries. Let's start from winery number one. So we are using just one feature, alcohol. And if you remember, we had 43 bottles from this winery. So, we have 43 numbers, 43 alcohol levels. And what I'm showing over here is just a histogram plot of these numbers. **As you can see it's very jagged. As a histogram always is, if you have very few data points. So how do we fit a Gaussian to this? Well, to fit a Gaussian, we just need to figure out the mean and the variance**. So, we take our 43 numbers, the 43 alcohol levels and we compute their mean and we compute the variance, that's very quick, it's just two lines of code. And what do we get? This is the distribution we get. The mean turns out to be 13.72. The standard deviation is 0.44. So sigma is 0.44, which means the variance is the square of that, sigma squared is roughly 0.2. And this is what the distribution looks like. This red line is the density that we fit for winery number one. As you can see, it's a nice, smooth distribution compared to the lumpy mess that we started with.
So, this will be done for winery one and we're just gonna go ahead and do exactly the same thing for winery two and winery three.

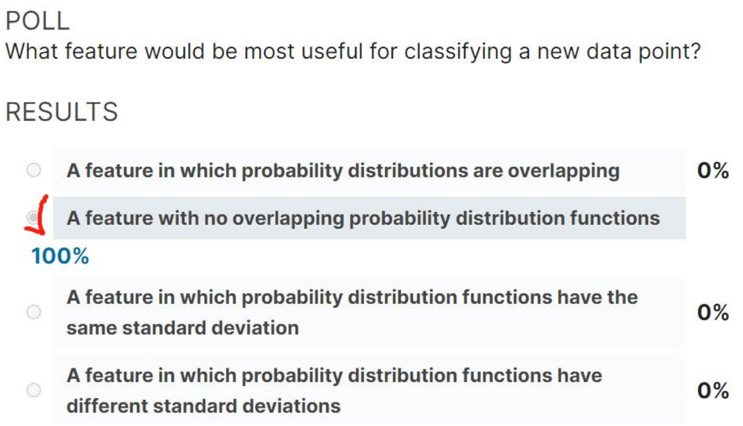
And this is the result. So, we already talked about winery one. For winery two, it turned out that I think we had 51 bottles, so 51 numbers. And their mean was 12.3 and the variance was 0.28. And for winery three, the mean was 13.2 and the variance was 0.27. And you can see the three curves over here. These are p one, p two, and p three. The distributions for each of the three classes. Now, are these distributions nicely separated from each other? No, not really. In fact, they're kind of on top of each other. And this really gives us a hint that using this one feature alone, we are not gonna be able to classify wineries very accurately. But let's go ahead and see what we get anyway.
So, given all these information, given this stuff and given these π's and these p one, p two, p three, how do we classify a new bottle of wine? So, we get a new bottle of wine, we measure its alcohol level, x, and then, we pick the answer j, one, two or three for which πj times pj of x is the highest.

- For example, let's say that our new bottle of winehas an alcohol level of 15, so it's over here. Which one are we gonna pick? Well, the black and green ones, that's wineries' two and three, have almost zero density at that point. So, we're gonna pick the red one, winery number one.

- What if the bottle of wine has an alcohol level of 11? Then we would pick winery number two, the black one.

- What if it has an alcohol level of something like 12.5, over here? What will we pick? Well, we wouldn't pick the red one. We'd pick either the black or the green. So, this is a case, in which winery two and winery three, those two distributions, p two and p three, have roughly the same density. So, which one would we pick? Well, we would pick winery number two because it has the higher class weight. So, remember, we are multiplying these two numbers together. So, in that case, if the alcohol level is something like 12.7, we would end up picking winery number two. So, we do all this and what kind of performance do we get? It turns out that the test errors, we had a test set of 48 points. This misclassifies 14 of them and so the test error is 29%. That's not very good, but it's way better than random.

Well, so we have seen our first classifier using the generative approach. It's performance was somewhat mediocre because it was based on a single feature. When we throw in a lot more features, we'll be able to improve its performance dramatically. But in order to see how to be able to do this, we are going to delve a little further into probability.



POLL
What feature would be most useful for classifying a new data point?

RESULTS

○ A feature in which probability distributions are overlapping 0%

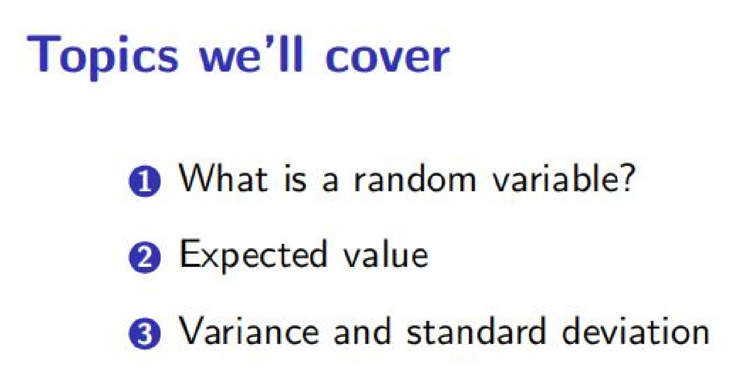
[✓ A feature with no overlapping probability distribution functions] 100%

○ A feature in which probability distribution functions have the same standard deviation 0%

○ A feature in which probability distribution functions have different standard deviations 0%

总结：

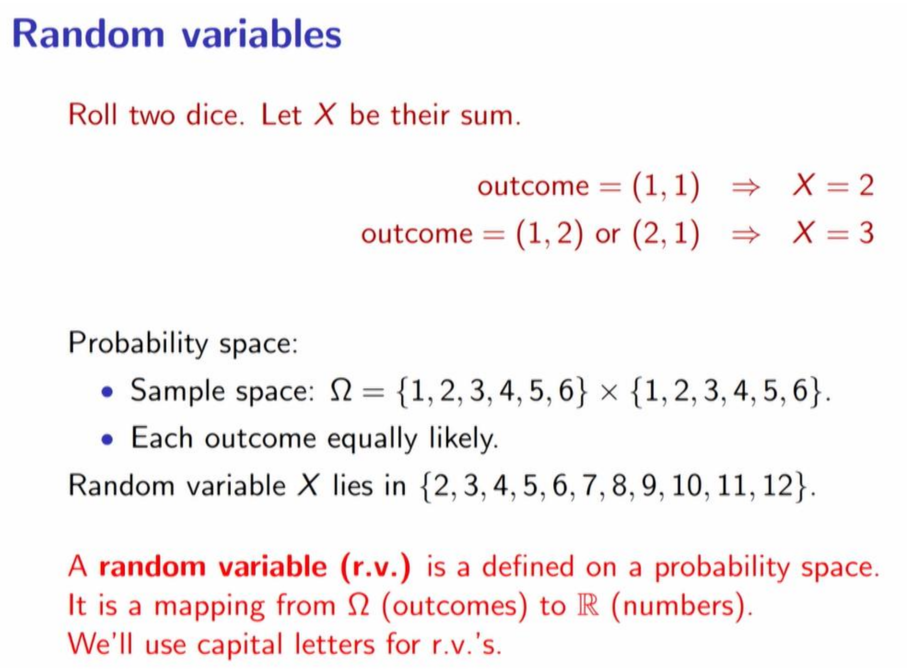
2.4 Probability Review II: Random Variables, Expectation, and Variance



Topics we'll cover

1. What is a random variable?
2. Expected value
3. Variance and standard deviation

In our first probability review, we talked about the notion of a *probability space*, so now we are ready to move on to the next major concept, which is that of a *random variable*.
What exactly is a random variable, and how do we compute their expected value, variance, and standard deviation?



Random variables

Roll two dice. Let X be their sum.

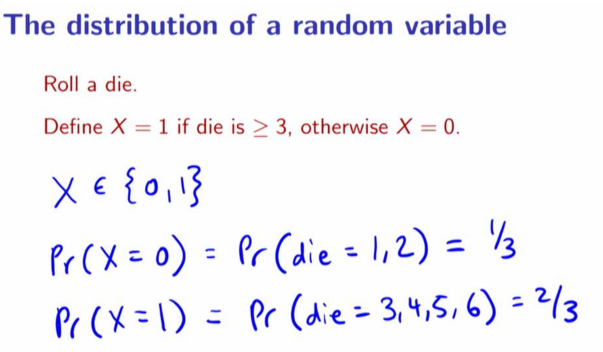
$$\text{outcome} = (1,1) \quad \Rightarrow \quad X = 2$$
$$\text{outcome} = (1,2) \text{ or } (2,1) \quad \Rightarrow \quad X = 3$$

Probability space:

- Sample space: $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$.
- Each outcome equally likely.

Random variable X lies in $\{2,3,4,5,6,7,8,9,10,11,12\}$.

A **random variable (r.v.)** is a defined on a probability space.
It is a mapping from Ω (outcomes) to $\mathbb{R}$ (numbers).
We'll use capital letters for r.v.'s.



The distribution of a random variable

Roll a die.

Define $X = 1$ if die is ≥ 3, otherwise $X = 0$.

$X \in \{0,1\}$

$Pr(X=0) = Pr(die = 1,2) = 1/3$

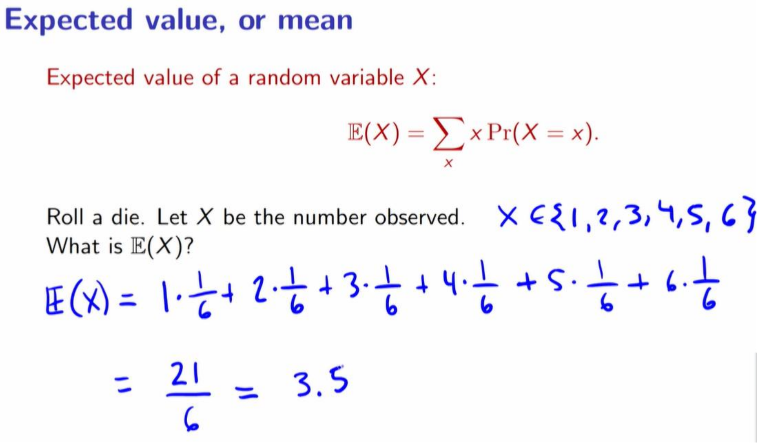
$Pr(X=1) = Pr(die = 3,4,5,6) = 2/3$

As usual, let's start with an example.Suppose you roll two dice, and you're interested in their sum. Let's call this X. So you aren't interested in exactly what die one comes out to, or die two comes out to, just in what their sum is, just in what they add up to. For example, if the first die is one and the second is one, then X equals two. If the first die is one and the second is two, then X equals three, and so on. This is a situation where the probability space has 36 different outcomes: six choices for the first die, and six choices for the second die. But we're only interested in a particular aspect of this outcome, the sum of the two die, and this aspect, which we call X, can only take on 11 possible values: two, three, all the way up to 12. X is a random variable.

In general, when we have a probability space, a random variable is any function of the outcome. It captures some aspect of the outcome that we care about. That's it.

Let's see how we can determine the distribution of a random variable from the information in the underlying probability space. Here's another simple example. We roll a die, and this time we don't care about exactly what the die comes out to. All we care about is whether or not it's greater than or equal to three. Given our interest, the way we define a random variable is we say: we have X, and we're only gonna have it take on two possible values. We'll say X is one if the die is greater than or equal to three, and otherwise X is zero. So the random variable we've defined in this case takes on values either zero or one, and let's go ahead and compute its distribution. The probability that X is zero is the probability that this die comes out to either one or two, which is 1/3, and the probability that X is one is simply the probability that the die comes out to three, four, five, or six, which is 2/3. In this way we can figure out the distribution of a random variable just by looking at the underlying probabilities of outcomes.

总结：



Expected value, or mean

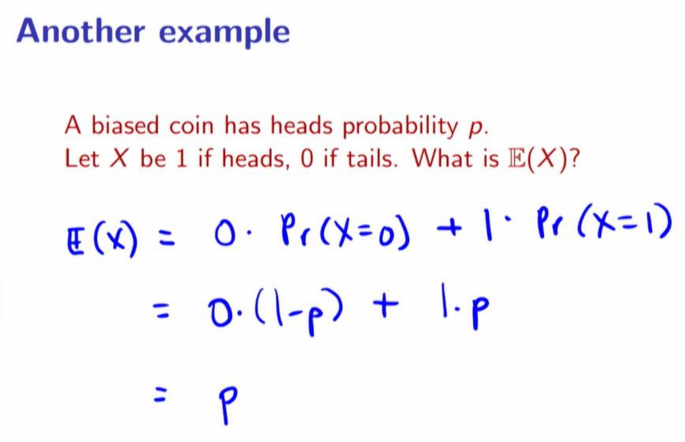
Expected value of a random variable X:

$$\mathbb{E}(X) = \sum_x x \Pr(X = x).$$

Roll a die. Let X be the number observed. What is $\mathbb{E}(X)$? $X \in \{1,2,3,4,5,6\}$

$$\mathbb{E}(X) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + 3\cdot\frac{1}{6} + 4\cdot\frac{1}{6} + 5\cdot\frac{1}{6} + 6\cdot\frac{1}{6}$$

$$= \frac{21}{6} = 3.5$$



Another example

A biased coin has heads probability p.
Let X be 1 if heads, 0 if tails. What is $\mathbb{E}(X)$?

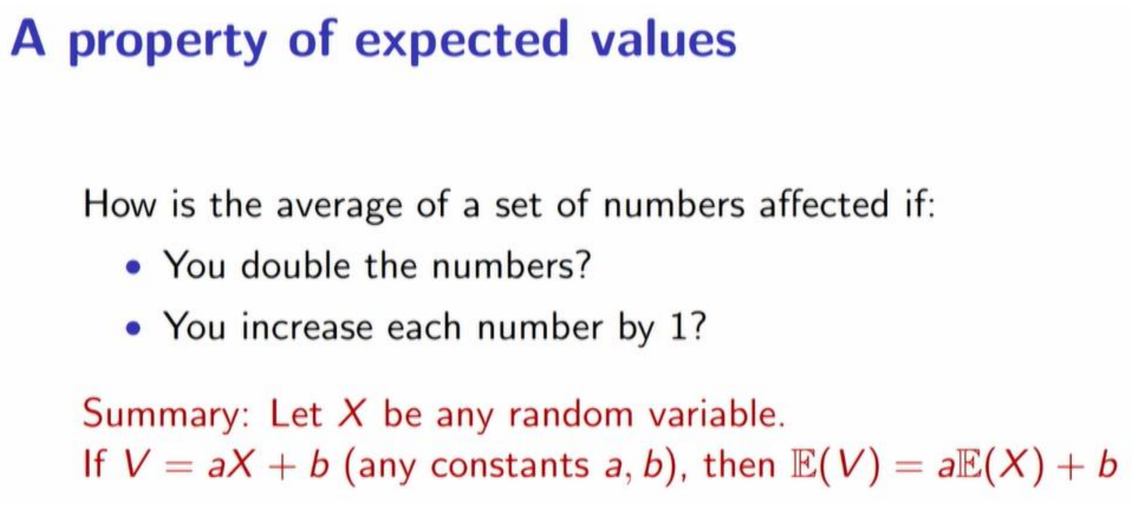
$$\mathbb{E}(X) = 0\cdot \Pr(X=0) + 1\cdot \Pr(X=1)$$

$$= 0\cdot(1-p) + 1\cdot p$$

$$= p$$

Once we have a random variable's distribution, the next order of business is usually to compute the expected value of the random variable, also known as its mean. The expected value of X is simply the average value of X if we were to repeat the experiment over and over and over again. It is a weighted average of all the possible values that X can take, where each value is weighted by its probability of occurring. It's a simple formula, and let's see it in use. Let's look at a simple example. We roll a die, let X be the number observed. What is its expected value? In this case, X is something that takes on values one, two, three, four, five, or six, and so the expected value of X, plugging into this formula, is simply an enumeration over all its possible values, so it takes on value one with probability 1/6, it takes on value two with probability 1/6, three with probability 1/6, four times 1/6, five times 1/6, and six times 1/6. So you get the sum of numbers from one to six, which is 21, so 21/6, which is 3.5, and so that's the expected value of X.

Let's do another example. This time you have a coin, but it's a biased coin, so the chance that it comes up heads is some number p, in the range zero to one. You flip this coin, and the random variable is the random variable X is one if the coin comes up heads, and zero if it comes up tails. What is the expected value of X? Once again, we just write out the formula. We look at all the possible values that X can take. It can take on the value zero, or it can take on the value one. The probability that it's zero is one minus p, and the probability that it's one is p, and so the expected value of X is p.



A property of expected values

How is the average of a set of numbers affected if:

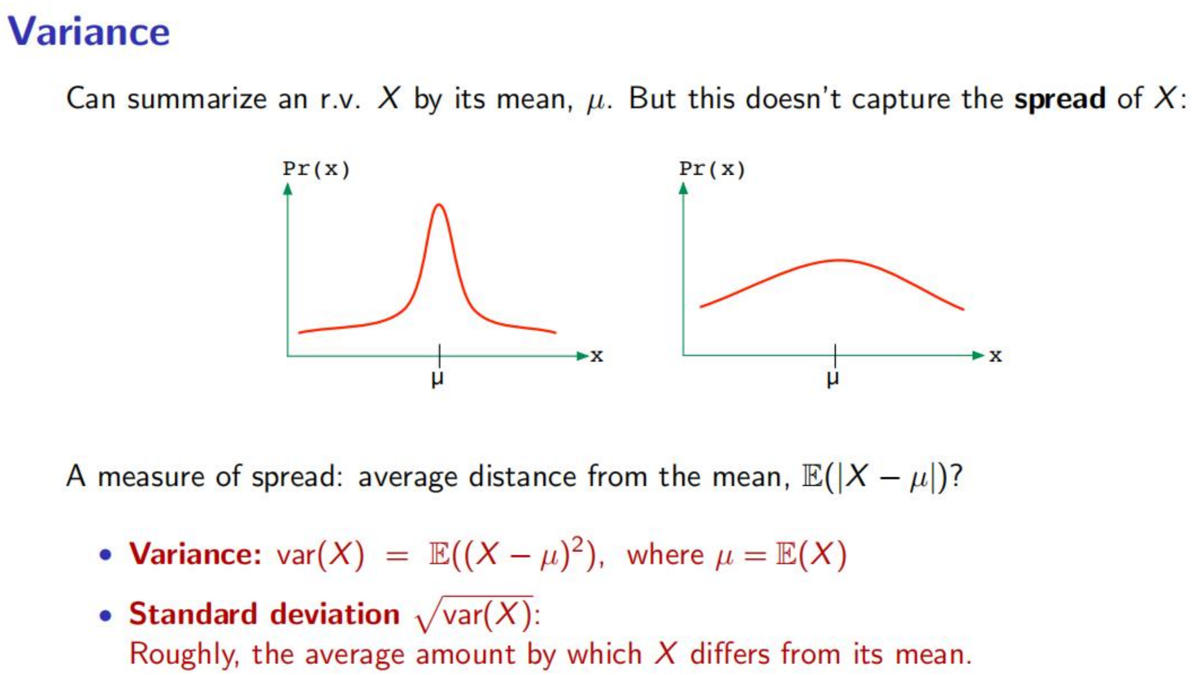
- You double the numbers?
- You increase each number by 1?

Summary: Let X be any random variable.
If $V = aX + b$ (any constants a, b), then $\mathbb{E}(V) = a\mathbb{E}(X) + b$

Now that we've some examples of expected values, let's look at a key property of these means. Suppose you have a bunch of numbers, and you double all the numbers. What happens to the average? Well, if you double all the numbers, the average also gets doubled. What if, instead of doubling the numbers, you increase every number by one. What happens to the average? So if every number increases by one, then the average also increases by one.

Let's summarize this in a slightly more general way. If you have any random variable X, and you define a new random variable V, to be some constant times X plus some other constant, like two X plus one, or four X minus three, what is the expected value of V? What is the mean of V? Can you figure it out from the mean of X? Yes, it's simply the mean of X times a plus b, and this is the linearity property that helps simplify many calculations.

总结：



Variance

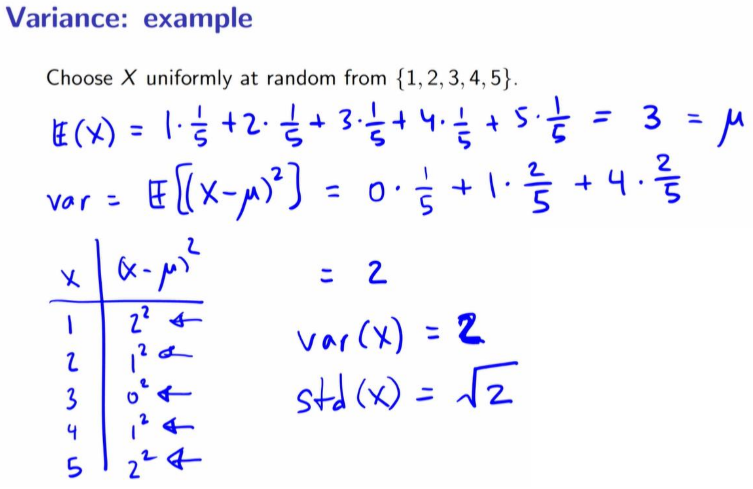
Can summarize an r.v. X by its mean, μ. But this doesn't capture the **spread** of X:

A measure of spread: average distance from the mean, $\mathbb{E}(|X - \mu|)$?

- **Variance:** $\text{var}(X) = \mathbb{E}((X-\mu)^2)$, where $\mu = \mathbb{E}(X)$
- **Standard deviation** $\sqrt{\text{var}(X)}$:
  Roughly, the average amount by which X differs from its mean.

In general, a random variable X can take on many possible values. What if we just wanted to summarize the distribution of X by a single number? What number would we pick? Probably its mean. Let's go ahead and denote the mean by the Greek letter mu. This is a nice summary. This is a concise and useful summary, but one of the things that it does not capture is the spread of X. Here's an example of two different distributions. They both have exactly the same mean mu, but the one on the left is much more tightly concentrated around its mean, while the one on the right is much more diffused. They have the same mean, but their spread is very different. So the mean isn't capturing certain very important properties of the distribution.
What if we were allowed a second number with which to summarize the distribution? Could we pick a number that somehow captures this spread? Here's an idea: why not use the average distance from the mean? So the expected value of the absolute difference between X and its mean, so in the distribution on the left, the average distance of X from its mean would probably be something like this, whereas the distribution on the right, the average distance of X from its mean might be something like this. This would actually work very well, and we could certainly use it, and in fact some people do use it, but it turns out that mathematically it's more convenient to look at the average square distance of X from the mean, and this is what we call the **variance** of X. The variance is the average square distance from the mean, and we'll see some properties of this in a second.

Now intuitively, what we do want is the typical distance from the mean, rather than the typical square distance, and so what we end up doing is to take the square root of the variance, and this is the **standard deviation** of X. The standard deviation behaves in the way we'd want it to. On the left, for example, the standard deviation would be something small, whereas on the right it's something larger.



Variance: example

Choose X uniformly at random from $\{1, 2, 3, 4, 5\}$.

$$\mathbb{E}(X) = 1\cdot\frac{1}{5} + 2\cdot\frac{1}{5} + 3\cdot\frac{1}{5} + 4\cdot\frac{1}{5} + 5\cdot\frac{1}{5} = 3 = \mu$$

$$\text{var} = \mathbb{E}[(X-\mu)^2] = 0\cdot\frac{1}{5} + 1\cdot\frac{2}{5} + 4\cdot\frac{2}{5}$$

$$= 2$$

x	$(x-\mu)^2$
1	2^2 ←
2	1^2 ←
3	0^2 ←
4	1^2 ←
5	2^2 ←

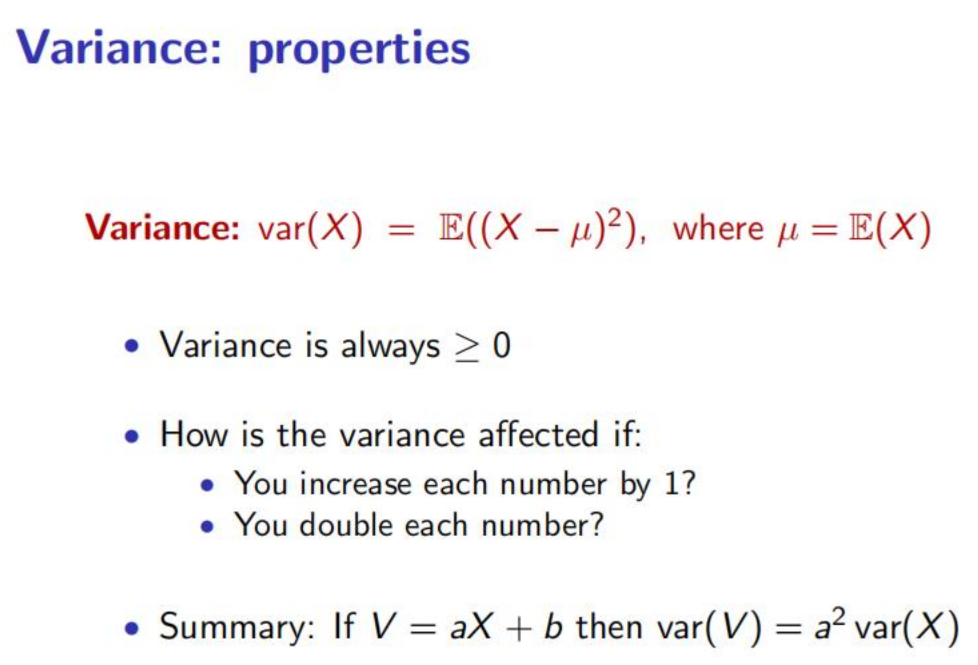
$$\text{var}(X) = 2$$

$$\text{std}(X) = \sqrt{2}$$

Let's see how we'd compute these quantities in a simple example. We pick a random variable X from one, two, three, four, five, all equally likely. What is the variance of X? The first thing we need to do is to compute the expected value of X, the mean, and in this case we'd expect it to just be the number in the middle, which is three. But let's double-check. We'll plug it into the formula. The expected value of X is enumerated over all possible outcomes. It's one with probability 1/5, two with probability 1/5, three times 1/5, four times 1/5, and five times 1/5. So you add up the numbers from one to five, so that is nine and three, 15/5, which is three, which is exactly what we expected. Good. Now we have the expected value of X. This is what we've been calling mu.

To compute the variance, we need the expected value of X minus mu squared. Let's look at the different values that X can take on, and X minus mu squared. When X is one, X minus mu squared, so mu is equal to three, is two squared. When X is two, X minus mu squared is one squared. When X is three, X minus mu squared is zero squared. Four, one squared. And five, two squared. So the expected value of X minus mu squared well, X minus mu squared can take on just three possible values: zero, one, or four. It takes on value zero with probability 1/5, that's this line of the table. It takes on value one with probability 2/5, that's these two lines of the table. And it takes on value four with probability 2/5, that's these two lines of the table. So this works out to two. So the variance of X is two, which means that the standard deviation of X is the square root of two. The square root of the variance.

总结：

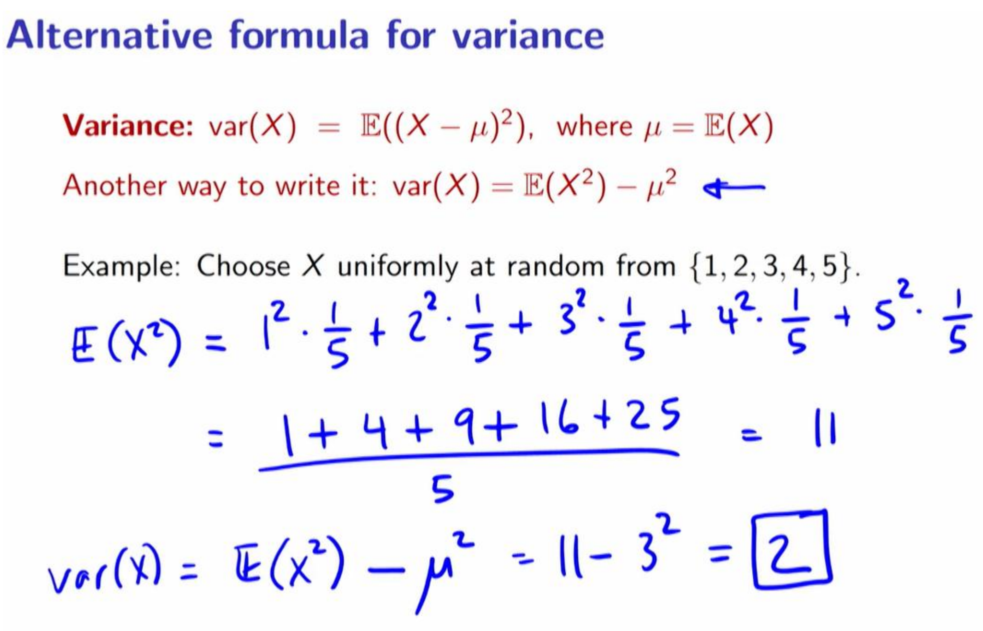


Variance: properties

Variance: $\mathrm{var}(X) = \mathbb{E}((X-\mu)^2)$, where $\mu = \mathbb{E}(X)$

- Variance is always ≥ 0
- How is the variance affected if:
  - You increase each number by 1?
  - You double each number?
- Summary: If $V = aX + b$ then $\mathrm{var}(V) = a^2\,\mathrm{var}(X)$

Let's look at some properties of the variance.
First of all, the variance is always a non-negative number. Why? Because it's the average square distance from the mean, and the average of positive numbers has to be positive. Then let's do the same thought experiment we did with averages. Suppose you have a bunch of numbers, and you increase each number by one. What happens to the variance? If you increase each number by one, everything shifts over a little bit, but the spread remains the same, so the standard deviation remains the same, and the variance remains the same, so the variance is unaffected if you shift numbers by one. What happens if you double each number? Well if you double the numbers, then the spread also gets doubled, so the standard deviation is doubled, which means that the variance, which is the square of the standard deviation gets multiplied by four. Let's summarize this by a general formula. If you have a random variable X, and then you define a new variable, which is something like two X plus one, or in general a X plus b, then the variance of this new variable is a squared times the variance of X.



Alternative formula for variance

Variance: $\mathrm{var}(X) = \mathbb{E}((X-\mu)^2)$, where $\mu = \mathbb{E}(X)$

Another way to write it: $\mathrm{var}(X) = \mathbb{E}(X^2) - \mu^2$ ←

Example: Choose X uniformly at random from $\{1, 2, 3, 4, 5\}$.

$$\mathbb{E}(X^2) = 1^2 \cdot \frac{1}{5} + 2^2 \cdot \frac{1}{5} + 3^2 \cdot \frac{1}{5} + 4^2 \cdot \frac{1}{5} + 5^2 \cdot \frac{1}{5}$$

$$= \frac{1+4+9+16+25}{5} = 11$$

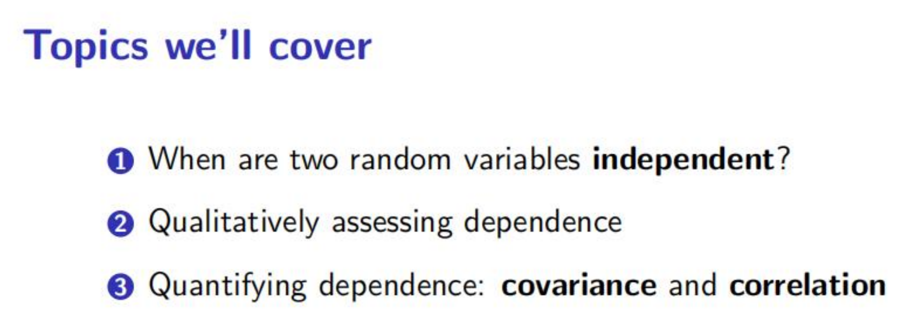
$$\mathrm{var}(X) = \mathbb{E}(X^2) - \mu^2 = 11 - 3^2 = \boxed{2}$$

There's one last property of variance that will be very useful, and that is that there's another formula for it. There's a different way to calculate it. It's this one over here. The variance of X, it turns out, is also the expected value of X squared minus the mean squared. It always works out to the same number. Let's see this at work, so let's go back to our earlier example, where we chose X uniformly from one, two, three, four, five. Now in this case, we saw that the variance was two, using the original formula. Let's try it again using the new formula. What is the expected value of X squared? X squared could be one squared with probability 1/5, or two squared with probability 1/5, or three squared with probability 1/5, or four squared with probability 1/5, or five squared with probability 1/5. What does this work out to? One plus four, plus nine, plus 16, plus 25, over 5, which is 55/5, which is 11. So the variance of X is the expected value of X squared minus mu squared. Now mu is the mean, we already know that's three. So it's 11 minus nine, which is 2, exactly as we got before.

We've seen some basic properties of random variables, and next time we'll get to the situation that we've gradually been building towards, a situation where we have multiple random variables: blood pressure, temperature, heart rate, disease, and we're interested in how they interact with each other. That's what we're gonna be needing to build generative models in higher dimension.

总结：

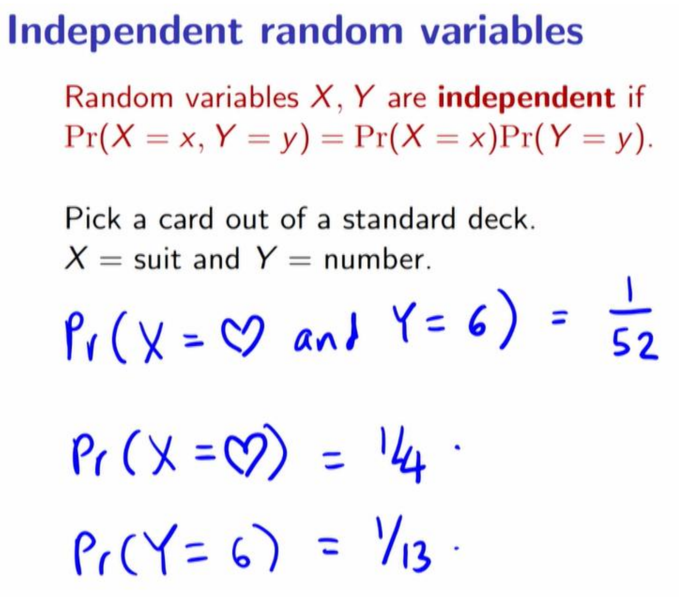
2.5 Probability Review III: Measuring Dependence



Topics we'll cover

1. When are two random variables **independent**?
2. Qualitatively assessing dependence
3. Quantifying dependence: **covariance** and **correlation**

When we're dealing with multiple random variables, for instance, a patient's heart rate, cholesterol level, blood pressure, whether they're ill, and so on, the easiest situation is when all these random variables are independent of each other, when they don't influence each other. In those cases, we can just model each of them separately, and not worry about interactions between them. The more interesting situation by far though is when they are dependent, and then we need ways of understanding and quantifying this dependence.



Independent random variables

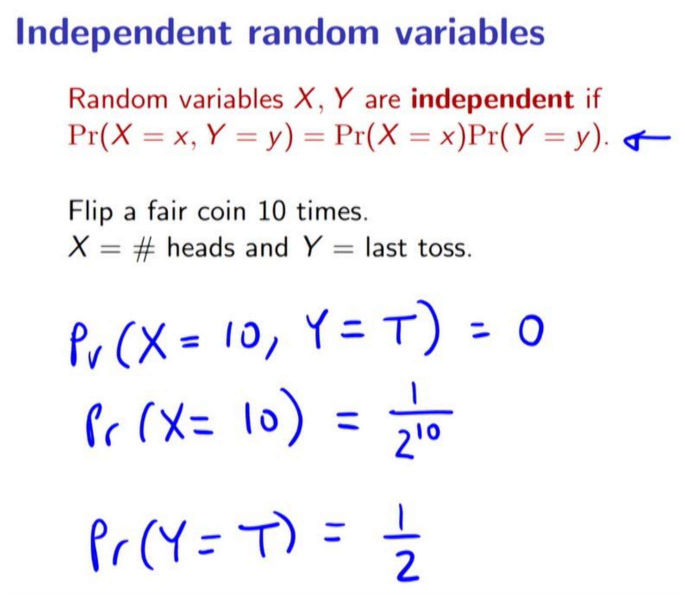
Random variables X, Y are **independent** if $\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$.

Pick a card out of a standard deck.
X = suit and Y = number.

$$\Pr(X = \heartsuit \text{ and } Y = 6) = \frac{1}{52}$$

$$\Pr(X = \heartsuit) = 1/4 .$$

$$\Pr(Y = 6) = 1/13 .$$



Independent random variables

Random variables X, Y are **independent** if $\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$. ←

Flip a fair coin 10 times.
$X = \#$ heads and Y = last toss.

$$\Pr(X = 10, Y = T) = 0$$

$$\Pr(X = 10) = \frac{1}{2^{10}}$$

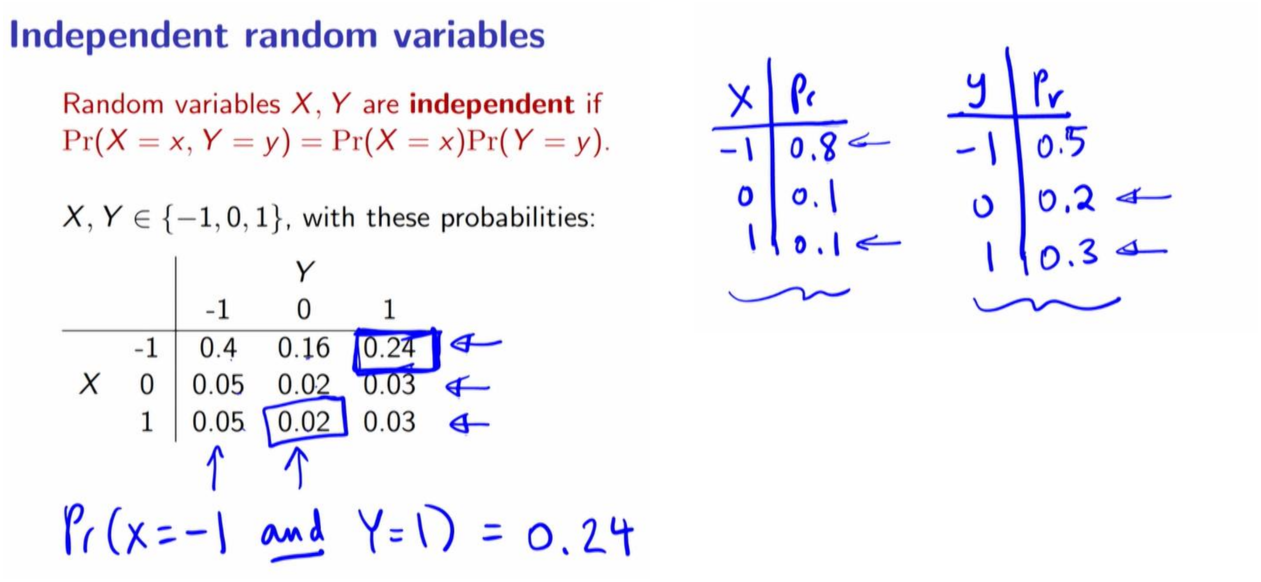
$$\Pr(Y = T) = \frac{1}{2}$$

So today, we'll begin by formally defining what it means for two variables to be independent, and then we'll start looking at dependence more closely. Okay, so this is the formal notion of independence. We say two random variables, X and Y are independent, if the probability that X takes on some value little X, and simultaneously, Y takes on some value little Y. It's just a probability that X is little X and that Y takes on the value little Y. In other words, X being equal to little x does not either make it more likely or less likely that Y will be equal to little y. They don't influence each other.

- Let's see an example of this. Suppose you have a standard deck of 52 cards, and you pick a card at random. And you define two random variables based on this card. X is the suit, clubs, spades, hearts, and so on. And Y is the number on the card. Six, seven, eight, jack, queen, king. Are X and Y independent? Yes they are. Let's see why that's the case. Intuitively it makes sense. But formally, let's look at an example. What is the probability, for example, that X is heart and Y is six? What is the probability of getting the six of hearts? Well, it's one out of 52 cards, so it's one over 52. Now, what is the probability, just by itself that X is hearts? If you pick a card at random, the probability of getting a heart is a quarter since there are four suits. And what's the probability that Y by itself is six? If you pick a card at random, the probability of getting a six is one out of 13. And indeed, one over 52 is equal to a quarter times one over 13. And so these two variables are independent.

- Let's do another example. Suppose you have a fair coin and you flip it 10 times. And you let X be the total number of heads you see, and let Y be the very last toss, heads or tails. Are these independent of each other? So intuitively, these do not seem independent. For example, if X is large, if there were lots of heads, then it makes it more likely that the last toss was also a heads. It seems like they really influence each other. But let's see more precisely why this formula breaks down. Why is this formula violated? Let's just look at a particular setting. What is the probability that X is 10 and Y is tails? What is the probability that all 10 tosses are heads, but the last toss is a tails? Clearly zero, it cannot happen. But now if you look at these two events separately, the probability that all 10 coin tosses are heads is greater than zero, it's a half, times a half, times a half, 10 times. So that's one over two to the 10th. And the probability that the last toss is tails is just a half. And clearly zero is not equal to one over two to the 10 times a half. And so these variables are not independent. They are dependent.

总结：



Independent random variables

Random variables X, Y are **independent** if $\Pr(X=x, Y=y) = \Pr(X=x)\Pr(Y=y)$.

$X, Y \in \{-1, 0, 1\}$, with these probabilities:

		Y		
		-1	0	1
X	-1	0.4	0.16	0.24
	0	0.05	0.02	0.03
	1	0.05	0.02	0.03

$\Pr(x=-1 \text{ and } Y=1) = 0.24$

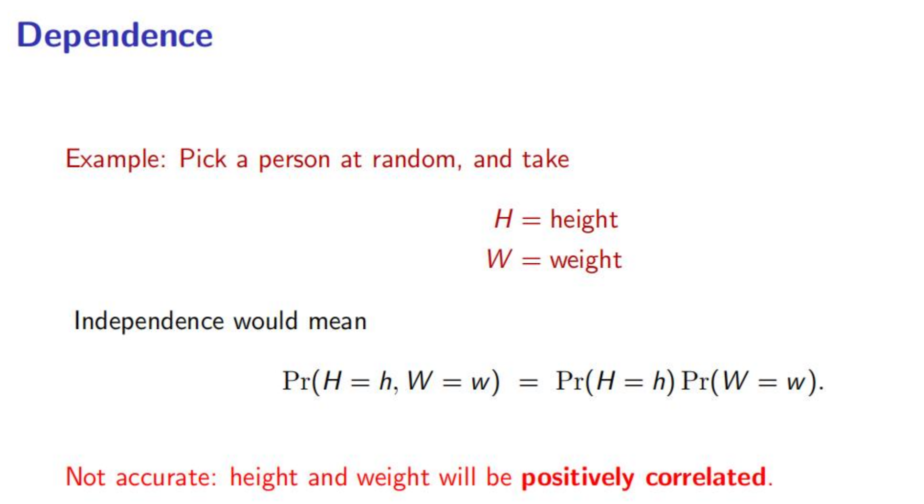
X	Pr
-1	0.8
0	0.1
1	0.1

y	Pr
-1	0.5
0	0.2
1	0.3

Here's another example. So now we have two variables, X and Y, which each take on values minus one, zero, or plus one. The distribution is summarized in this table over here. For example, this entry over here means that the probability that X is negative one and Y is one is a .24. So that's what this table means. Needless to say, the numbers in the table add up to one. Are X and Y independent? Well, we have to check this formula. In order to do that, we have everything we need for the left hand side, but now for the right hand side, we need the individual distributions for X just by itself, and Y just by itself. Let's obtain those.

- Let's look at X by itself. X can take on value either minus one, zero, or plus one. What are the probabilities of these? Well, the probability that X is minus one, X can be minus one in three ways. Any of the three situations on this line, so we just add them up. The probability that X is minus one is .4 plus .16 plus .24, which is .8. What is the probability that X is zero? That's on this line, so we add up those numbers, and we get .1. And what is the probability that X is one? That's this line, and that's again .1. And of course, these things add up to one again. That's the distribution of X.
- Now let's go ahead and do the same thing for Y. So we have Y over here. It takes on values minus one, zero, or plus one. And the probability that Y is minus one is this line over here, so it's .5. The probability that it's zero is this line here, so it's .2. And the rest is of course .3. So that's the distribution of Y.

Now to check if these are independent, we have to check that every entry in this table of the joint distribution, is the product of the corresponding entries in the individual distributions. So for example, let's look at this entry here. For X equals minus one, and Y equals one. Is this equal to this number times this number? Yes it is. That entry is okay, and in that way we keep going and we check all nine of these entries, and it turns out that in every case, the entry in this table is equal to the productof the corresponding entries in the individual tables. So this, for example, is equal to this times this. And so X and Y are independent.



Dependence

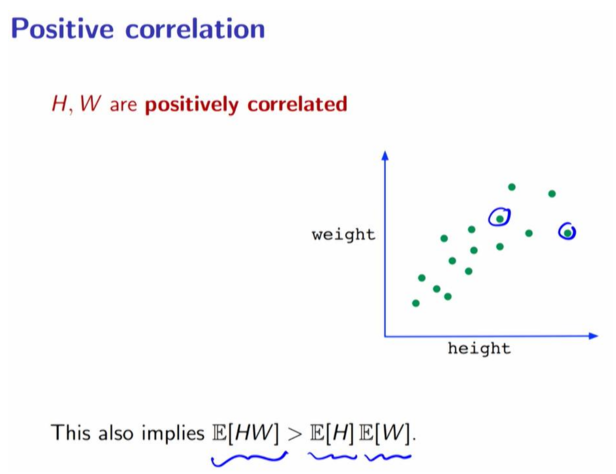
Example: Pick a person at random, and take

H = height
W = weight

Independence would mean

$$\Pr(H=h, W=w) = \Pr(H=h)\Pr(W=w).$$

Not accurate: height and weight will be **positively correlated**.



Positive correlation

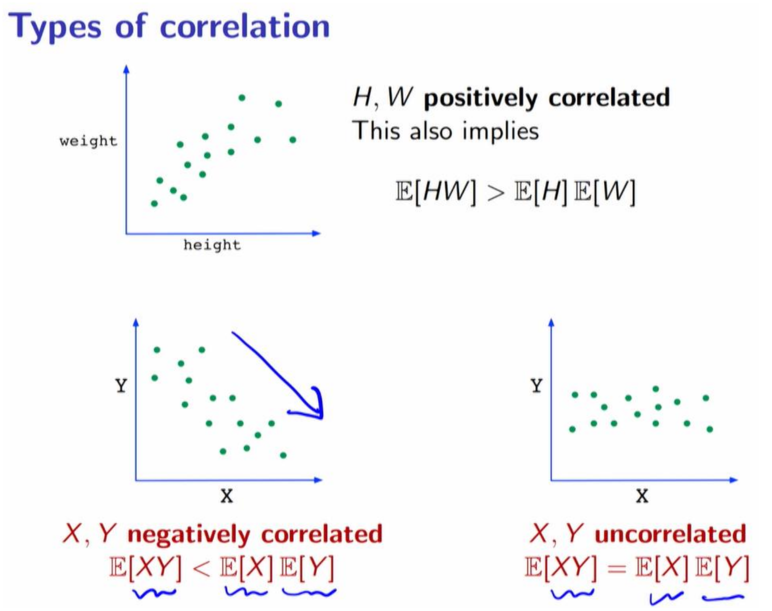
H, W are **positively correlated**

This also implies $\mathbb{E}[HW] > \mathbb{E}[H]\,\mathbb{E}[W]$.

So now we'll start talking about dependence. Let's just start with a little bit of a toy experiment. Suppose you pick a person at random from whatever city you're living in, and let H be their height and W be their weight. Are H and W independent of each other? Intuitively, one would think that they are not. The reason for that is that people who are taller are also likely to be heavier. So it seems like there's some sort of correlation between that. A positive correlation, in fact. We would expect height and weight to be positively correlated.

Here's a picture of what it might look like. What I've drawn over here is, suppose one took a sample of about 20 people and plotted their heights and weights. It might look something like this. There's a general upward trend. People who are taller tend to be heavier. Now, of course, knowing someone's height, you can't exactly predict their weight. And nor is it a hard and fast rule. For example, this person over here is taller than that person, and is also lighter. But there is a general trend that the points are sloping upwards. And what this also means is that we would expect the average of height times weight to be greater than the average height times the average weight. This is because the large values of H tend to be paired with the large values of W. So if you multiply them together, you get a sort of squaring effect. So the average of H times W is likely gonna be higher than the average H times the average W. This is what positive correlation looks like.

总结：



Types of correlation

H, W **positively correlated**

This also implies

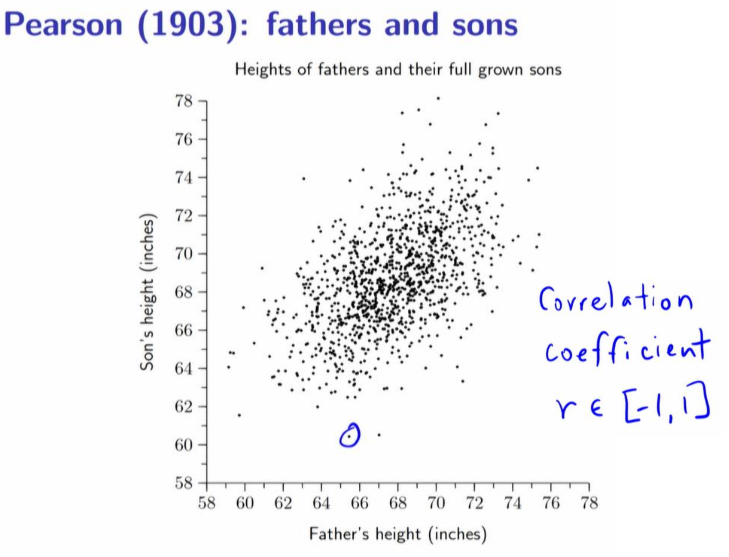
$$\mathbb{E}[HW] > \mathbb{E}[H]\,\mathbb{E}[W]$$

X, Y **negatively correlated**

$\mathbb{E}[XY] < \mathbb{E}[X]\,\mathbb{E}[Y]$

X, Y **uncorrelated**

$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$



Pearson (1903): fathers and sons

Heights of fathers and their full grown sons

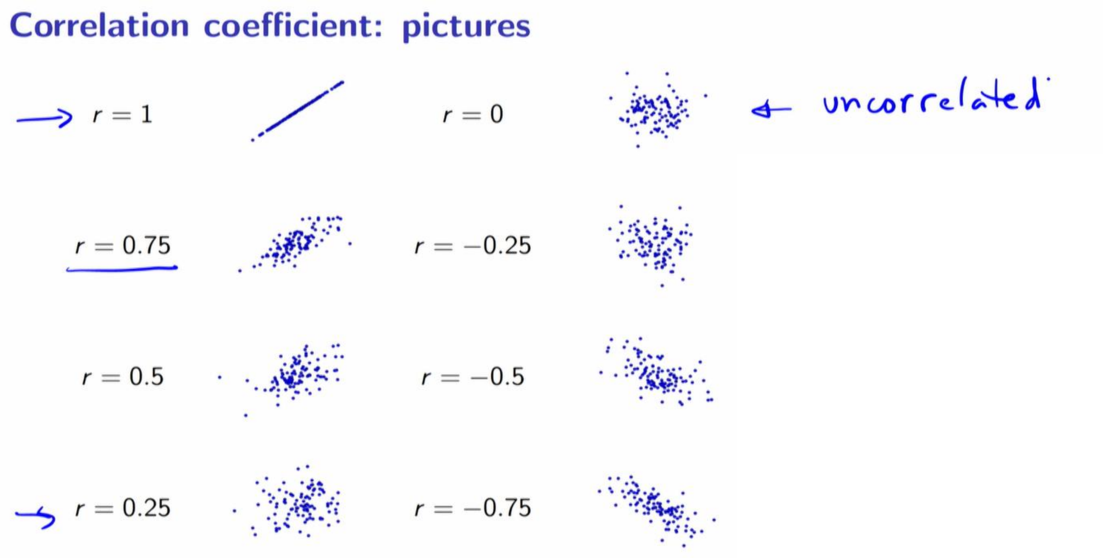
Correlation coefficient

$r \in [-1, 1]$

The other two types of correlationare negative correlation and a complete lack of correlation. In negative correlation, the points rather than seeming to slope upwards, are sloping downwards. Like this. And in this picture, the larger values of X tend to get paired with the smaller values of Y. As a result, one would expect that the average value of X times Y is actually less than the average X times the average Y.

And then there's a situation where no matter how large or small X is, the Ys on average tend to have the same mean. And that situation is called **uncorrelated**. Over there we would expect that the average value of X times Y is just the average X times the average Y. This is a spread of points that's neither sloping upwards nor downwards.

This is a very hand wavy and qualitative viewof correlation. Can we come up with a precise formula for this? An actual number that captures correlation? And this was something that was done by Carl Pearson at the beginning of the 20th century. He was interested in the correlation between a father's height and his son's height. And so what he did, is he obtained data from about 2,000 father/son pairs, and he plotted them. That's the scatter plot you see over here. On the X axis are the fathers' heights, and on the Y axis are the sons' heights. For example, this over here comes from a father/son pair in which the father is about 66 inches, and the son is about 60 inches. What you can see over here is a general upward trend. There does seem to be a positive correlation. What Pearson did, was to define something called a **correlation coefficient**. This is a number between minus one and plus onethat captured the degree of correlation. So a value of minus one means perfect negative correlation. And a value of plus one means perfect positive correlation.



Correlation coefficient: pictures

$r = 1$ $r = 0$ ← uncorrelated

$r = 0.75$ $r = -0.25$

$r = 0.5$ $r = -0.5$

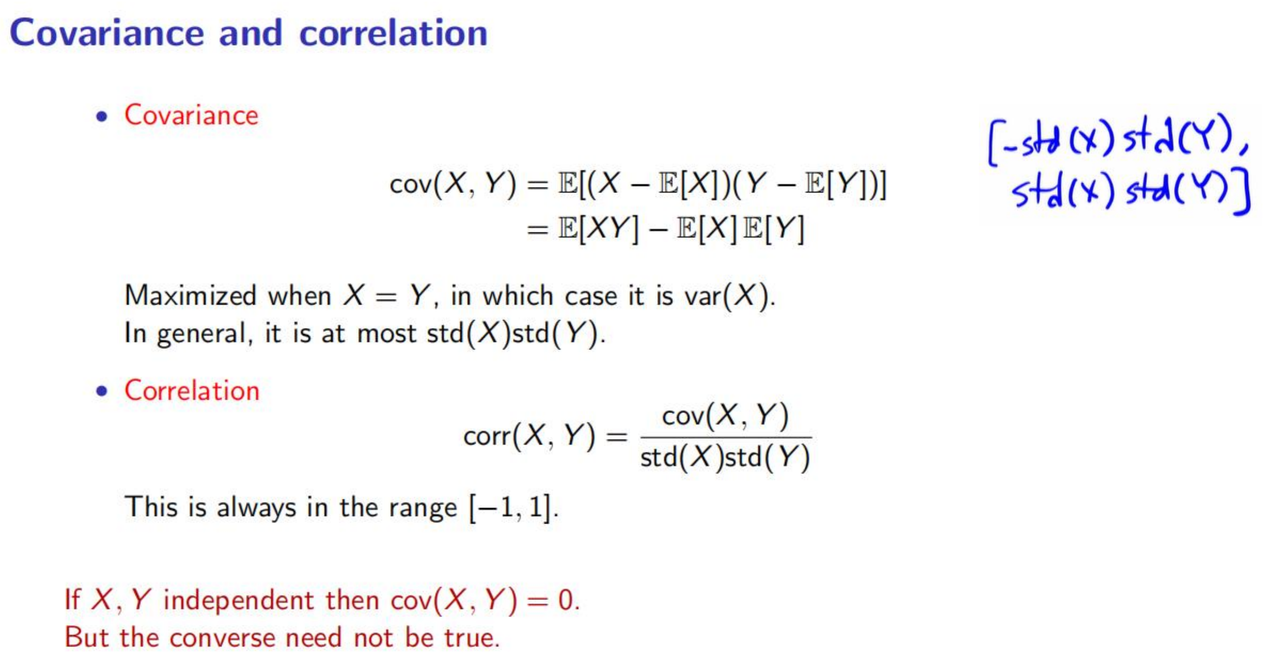
$r = 0.25$ $r = -0.75$

Let's see some examples of these values.

- When the correlation coefficient is 1, like in this picture, there's a perfect linear relationship between the two variables.
- Then when the correlation drops a little bit, to say .75, there's still a very clear upward trend.
- When the correlation drops further to .25, the points start to look a little bit more like a blob.
- The case where r equals zero, is the **uncorrelated** situation.
- Now when the values start to drop below zero, then we get the symmetric situation with negative correlation. It moves more and more towards a perfect linear, but negatively sloped relationship.

What is the formula? How does one obtain this value, r? In many ways, we anticipated it when we were back here (本页第一幅图). We said that one way to determine correlation is to compare the expected value of XY with the expected value of X times the expected value of Y. So let's just look at these two quantities, the left hand side and right hand side, and look at their difference. This is what we call the **covariance**.

总结：



Covariance and correlation

- Covariance

$$\begin{aligned}\text{cov}(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\ &= \mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

$[-\text{std}(X)\,\text{std}(Y),\ \text{std}(X)\,\text{std}(Y)]$

Maximized when $X = Y$, in which case it is var(X).
In general, it is at most std(X)std(Y).

- Correlation

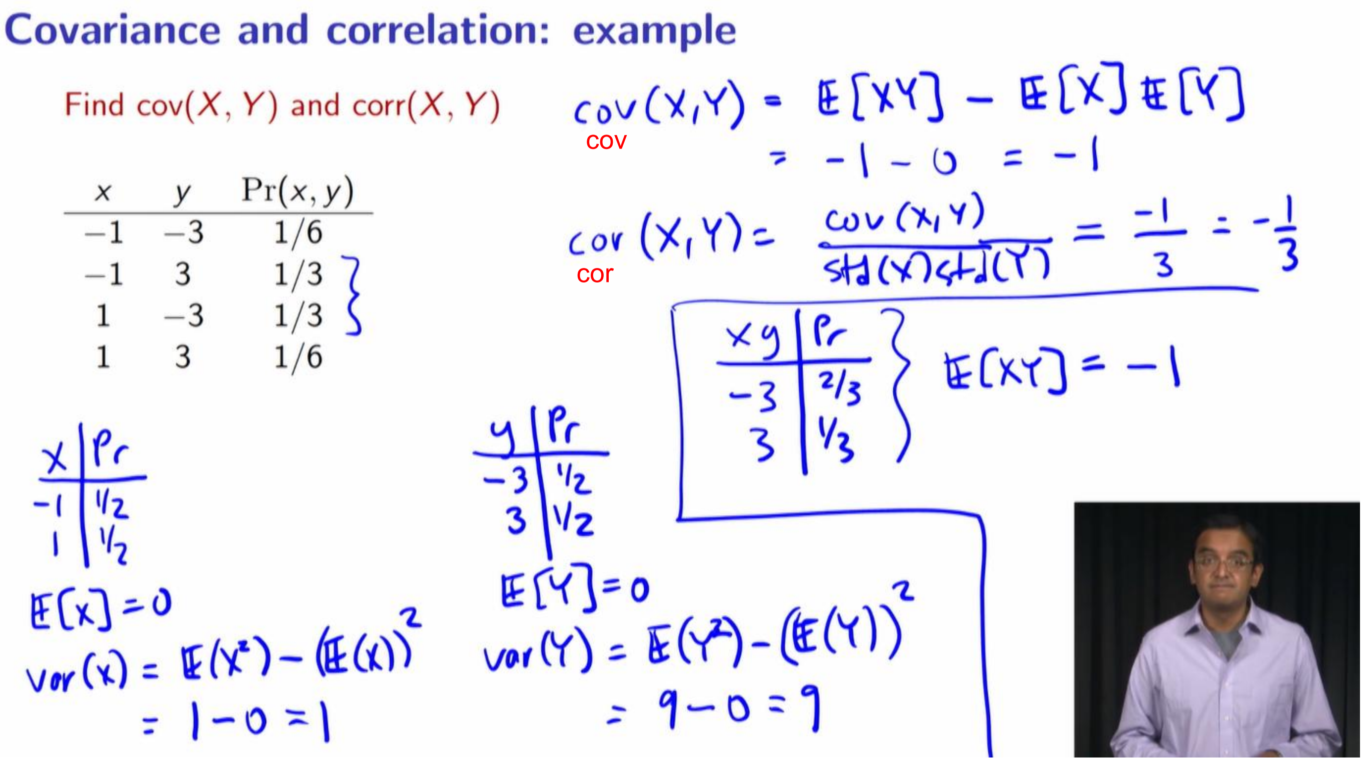
$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{\text{std}(X)\text{std}(Y)}$$

This is always in the range $[-1,1]$.

If X, Y independent then cov$(X,Y) = 0$.
But the converse need not be true.

The covariance between two variables, X and Y, is just the expected value of XY minus expected value of X times expected value of Y. When this is positive, it means we have a positive correlation. When it's negative, we have a negative correlation. And when it's zero, X and Y are uncorrelated. A very nice, simple measure of dependence. Now the one drawback to this, is that it actually is not necessarily in the range minus one to plus one. You can show mathematically that the range that covariance lies in is minus the standard deviation of X times the standard deviation of Y. That's the smallest it can be. All the way to the standard deviation of X times the standard deviation of Y. It lies somewhere in that range. If we want to normalize it, to a value in the range minus one to plus one, what should we do? Let's just divide it by the standard deviations of X and Y. And that is **Pearson's correlation coefficient**. It's a nice, normalized value in the range minus one to plus one.

One interesting question is how this relates to independence. When two variables are independent, we know that they are gonna be uncorrelated. But it turns out that the reverse need not be the case. Variables being uncorrelated is actually a much weaker condition than them being independent. It's quite possible to have two variables that have zero correlation, but are dependent on each other. So it's a one way relationship. **Independent means uncorrelated, but uncorrelated does not mean independent**.



Covariance and correlation: example

Find cov(X,Y) and corr(X,Y)

x	y	$\Pr(x,y)$
-1	-3	$1/6$
-1	3	$1/3$
1	-3	$1/3$
1	3	$1/6$

$$\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = -1 - 0 = -1$$

cov

$$\text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\text{std}(X)\text{std}(Y)} = \frac{-1}{3} = -\frac{1}{3}$$

cor

xy	Pr
-3	$2/3$
3	$1/3$

$\mathbb{E}[XY] = -1$

x	Pr
-1	$1/2$
1	$1/2$

$\mathbb{E}[X] = 0$

$$\begin{aligned}\text{var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= 1 - 0 = 1\end{aligned}$$

y	Pr
-3	$1/2$
3	$1/2$

$\mathbb{E}[Y] = 0$

$$\begin{aligned}\text{var}(Y) &= \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 \\ &= 9 - 0 = 9\end{aligned}$$

总结:

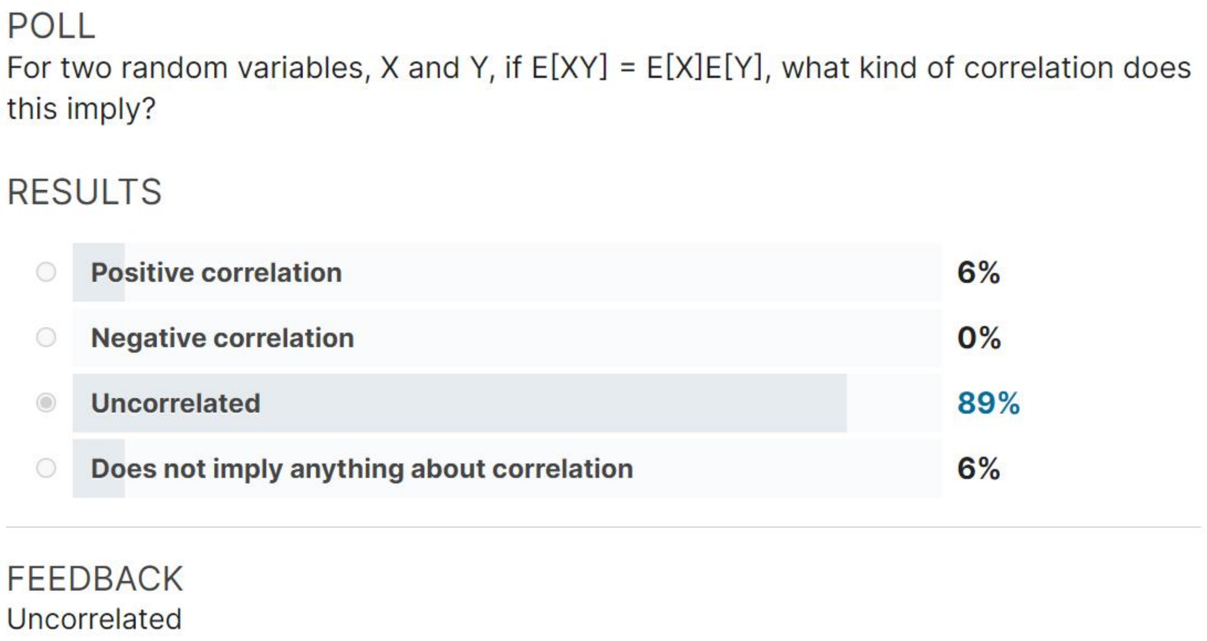
Let's finish with an example, a numeric example where we can actually see these numbers at work. Here we have a pair of variables, X and Y, and we want to figure out the covariance and correlation between them. We want to figure out the level of dependence between these two variables. Let's write down the formula again, just to remind ourselves. The covariance of X and Y is the expected value of XY minus the expected value of X times the expected value of Y. In order to figure this out, we do need to figure out the expected value of X and the expected value of Y. So let's start by doing that.

- Let's just look at X by itself. X takes on values either minus one or plus one. The probability that X is minus one, well X is minus one in these two lines, so the probability is a sixth plus a third, which is a half. So the probability that X is plus one, is also a half. And this means that the expected value of X is just zero. While we're at it, let's also do the variance. The variance of X is the expected value of X squared minus the expected value of X squared. Let's see what that is. X is either minus one or plus one, so X squared is always one. The expected value of X squared is just one. This we know is zero, so the variance is one. That means that the standard deviation of X is also one. The typical distance of X from the mean is one. And in fact, that makes perfect sense, since the distance of X from the mean is always exactly one.

- Now let's go ahead and do Y. Y is either minus three or plus three. And it's minus three in these two lines, and that's probability of a half. And plus three is probability of half, so the expected value of Y is again zero. And the variance of Y is the expected value of Y squared minus the value of Y, the whole thing squared. So Y squared, Y is either minus three or plus three. So Y squared is exactly nine. The expected value of Y squared is nine. The variance of Y is nine, which means its standard deviation is three. The typical distance of Y from its mean is three. In fact, it's always distance three from its mean. Great.

- Now in order to compute the covariance, we also need X times Y. Let's do that, too. Let's leave a little space for it here. X times Y is either minus three or plus three. The probability of minus three, it's minus three in these two lines. It's minus three probability 2/3 and plus three probability 1/3, which means that the expected value of XY is minus three times 2/3 plus three times 1/3, which is minus one. Good.

So now we have everything we need for the covariance of XY. It is minus one minus zero, so it's minus one. There is a negative correlation between X and Y.
Now to get the actual correlation coefficient, we divide by the standard deviations. We take the covariance and we divide it by the standard deviation of X and the standard deviation of Y. And so that's minus one. The standard deviation of X is one. The standard deviation of Y is three. So the correlation is -1/3. There's a weak negative correlation between these two variables.



POLL

For two random variables, X and Y, if E[XY] = E[X]E[Y], what kind of correlation does this imply?

RESULTS

○ Positive correlation 6%

○ Negative correlation 0%

● Uncorrelated 89%

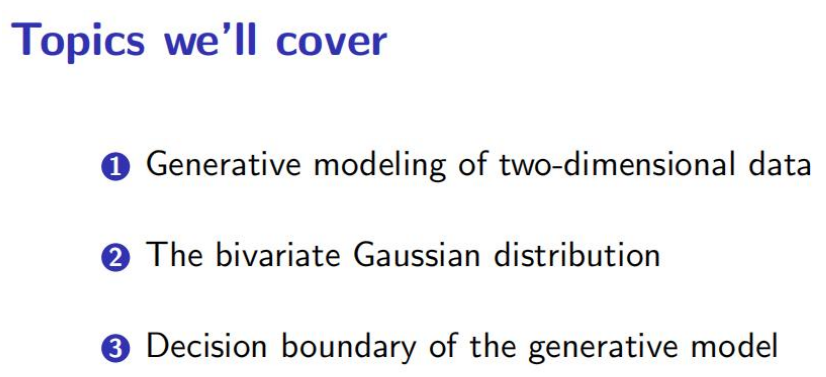
○ Does not imply anything about correlation 6%

FEEDBACK

Uncorrelated

总结：

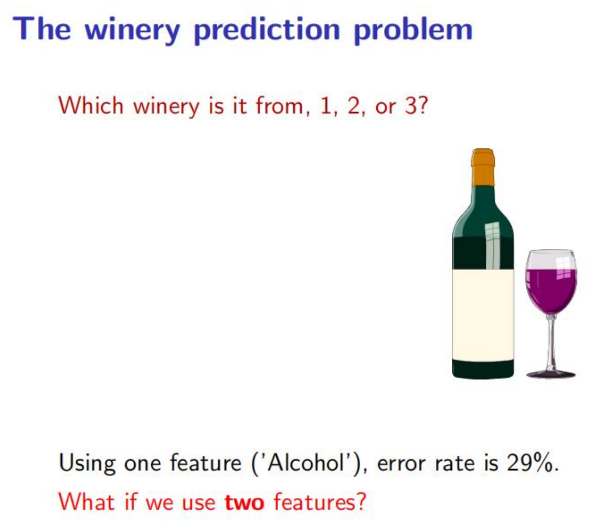
2.6 Two Dimensional Generative Modeling with the Bivariate Gaussian



Topics we'll cover

1. Generative modeling of two-dimensional data
2. The bivariate Gaussian distribution
3. Decision boundary of the generative model

We now continue our development of the generative approach to classification. So last time, we looked at the one-dimensional Gaussian distribution, and we saw how to build a classifier from it using the generative approach.
Today, we'll look at the Gaussian in two dimensions. One of the exciting things about this distribution is that it allows us to model the dependence between features. So, for example, if we are trying to predict whether somebody's ill based on their cholesterol level and their blood pressure, we don't have to pretend that these two features, cholesterol level and blood pressure, are independent. We can actually model the dependence between them in order to get a more accurate prediction.

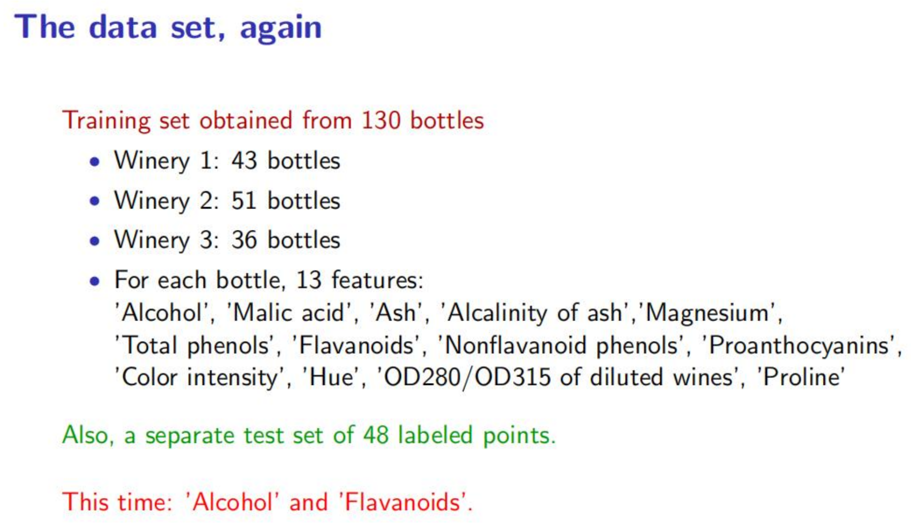


The winery prediction problem

Which winery is it from, 1, 2, or 3?

Using one feature ('Alcohol'), error rate is 29%.

What if we use **two** features?



The data set, again

Training set obtained from 130 bottles

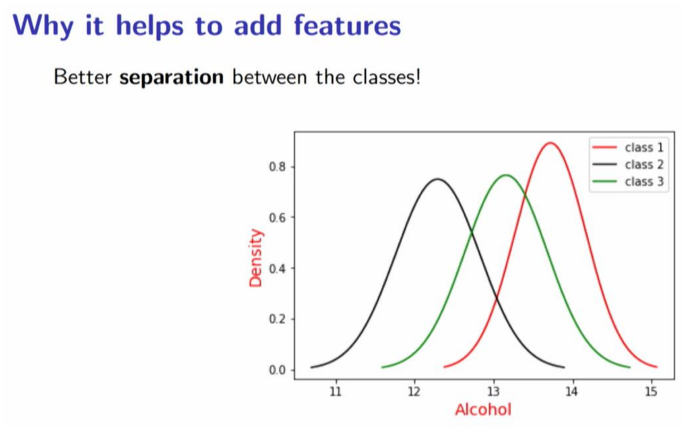
- Winery 1: 43 bottles
- Winery 2: 51 bottles
- Winery 3: 36 bottles
- For each bottle, 13 features:
  'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash','Magnesium', 'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins', 'Color intensity', 'Hue', 'OD280/OD315 of diluted wines', 'Proline'

Also, a separate test set of 48 labeled points.

This time: 'Alcohol' and 'Flavanoids'.

So our running example is the winery prediction problem. And if you recall, the situation here is that we have a bottle of wine. We have bottles of wine, and from each bottle, we measure 13 visual and chemical features. And we want to use these to predict which winery this bottle came from, winery one, two or three. Last time, we used just one out of the 13 features, and we found that using that one feature alone, we could get an error rate of about 29%. Today, we'll use two features, two out of the 13, and we'll see that the error rate drops quite dramatically.

So just to remind you about this data set, we had a training set obtained from 130 bottles, so 130 data points. 43 of them were from winery one, 51 from winery two and 36 from winery three. And from each bottle, we extracted 13 features that makes sense to wine experts. There was also a separate test set with 48 labeled points. Now last time, the feature that we used was alcohol. This time we'll use a second feature, flavonoids. These are a family of chemical compounds that show up in wine. Okay, so these are the two features we're gonna use.



Why it helps to add features

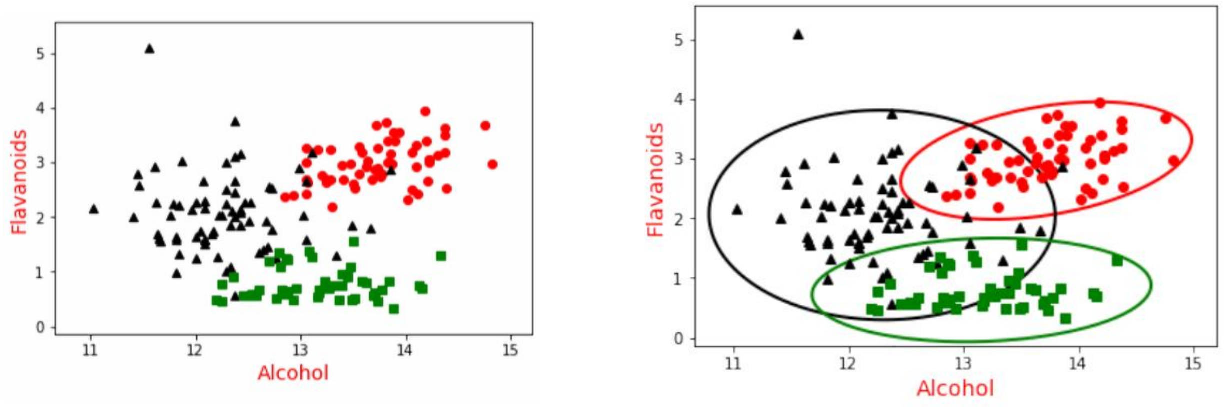
Better **separation** between the classes!

And let's just sort of think at a high level first about why it might be helpful to throw in a second feature.
So what happened last time is we just had this one feature, alcohol. And we fit a Gaussian to the alcohol levels from winery one, another Gaussian to the alcohol levels from winery two and a third one to the alcohol levels from winery three. Looking at the three Gaussians, it turned out that they aren't really very well separated from each other. In fact, they're kind of on top of each other. And this is why this one feature alone turned out to not be a great basis for doing classification.

总结：

Why it helps to add features

Better **separation** between the classes!

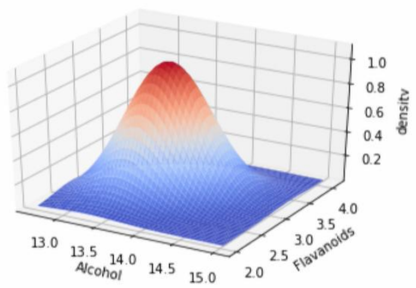


Error rate drops from 29% to 8%.

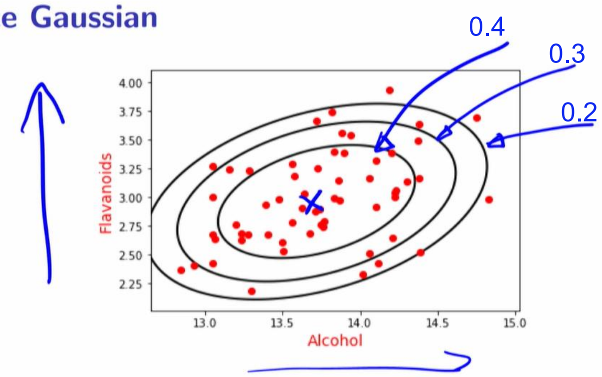
So let's see what happens when we throw in a second feature. So what I've shown here is a scatterplot. Each point is one of the 130 training points. The red ones are from winery one. The black ones are from winery two. The green ones are from winery three. And for each of the training points, I've shown two of the features, alcohol level and flavonoids, okay. So each point has been plotted in two dimensions. And you can see that there is actually quite a bit of separation between them. So this makes us hopeful. It gives us hope that these two features might be the basis for better classification. So what we're gonna do is to follow the generative approach and to fit a distribution to the data from winery one, a separate distribution to winery two and a separate distribution to winery three. And the distributions are gonna turn out to be something like this, okay. This might seem mysterious. What are these ellipses? So I will explain in a second. But roughly, this is the picture. And it turns out when you do this, the error rate drops quite substantially, from 29% to just 8%, okay.

So all this is great, but what are these distributions, and what are these ellipses? So let's hone in on one of them. Let's just look at the red one, which is from winery number one. Let's just look at those specific points.

The bivariate Gaussian



The bivariate Gaussian



Model class 1 by a bivariate Gaussian, parametrized by:

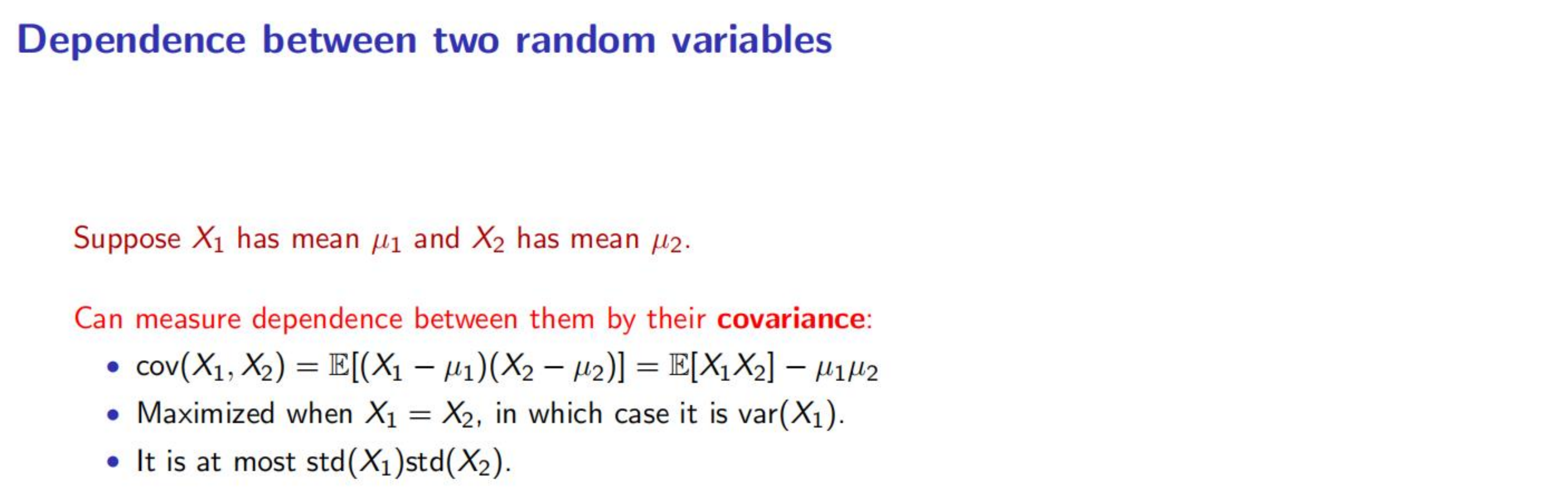
$$\underbrace{\text{mean } \mu = \begin{pmatrix} 13.7 \\ 3.0 \end{pmatrix}} \text{ and } \underbrace{\text{covariance matrix } \Sigma} = \begin{pmatrix} \boxed{0.20} & \overset{\text{covariance}}{\boxed{0.06}} \\ 0.06 & \underset{\text{variance}}{\boxed{0.12}} \end{pmatrix}$$

So when we fit a distribution to those, when we fit a Gaussian distribution to those 43 training points, this is the distribution we get, okay. So what is this thing? Well, so it's a distribution over the 2D plane, okay. So it's a distribution that assigns a value to every possible alcohol level and every possible flavonoid level. So at the base over there, you see the alcohol and flavonoid values. That's a 2D plane, and the height is the density at that point, okay. This is what a bivariate Gaussian distribution looks like.

Now for me, personally, it's a little hard to interpret these 3D pictures, so I much prefer to draw it like this, okay. So what I've drawn herein the red are the original training points, the 43 training points, the 43 two-dimensional points from winery one. Using these points, we fit a Gaussian distribution, and this distribution has two sets of parameters. It has a mean, and it has a covariance matrix. The mean is just a 2D point. It is the center of the distribution. It is the point of highest density. So we're fitting a distribution, and there's one point at which the density is highest. It's the center of the distribution, the mean. Now as you move away from the mean, the density drops off, and it drops off with this ellipsoidal contours. So let's say the density of the mean is 0.5. You can ask, "Let's look at all points "whose density is 0.4. "What are those points?" Well, it might be this ellipse, for example. And then you can say, "What are all points "with density 0.3?" Well, it might be this ellipse, which is concentric with the first one. And what are all

总结：

points with density 0.2? Well, it might be this ellipse, okay. So what I'm showing over here are just contour lines of this density. The point of the highest density is the mean, and then it falls off in these ellipsoidal contours.
So let's look at these parameters again. The mean makes a lot of sense. It's just the center of the density. What is this **covariance matrix**? So this is a two-by-two matrix. This number, over here, is the variance along the x one direction. It's just the variance of the alcohol value. So we have 43 numbers, 43 alcohol levels. We compute their variance, that's 0.2. This number, over here, is the variance along the second direction. It's just the variance of all the flavonoid numbers. The interesting thing is this number, over here. This is the covariance between the two features, between the alcohol level and the flavonoid level.



Dependence between two random variables

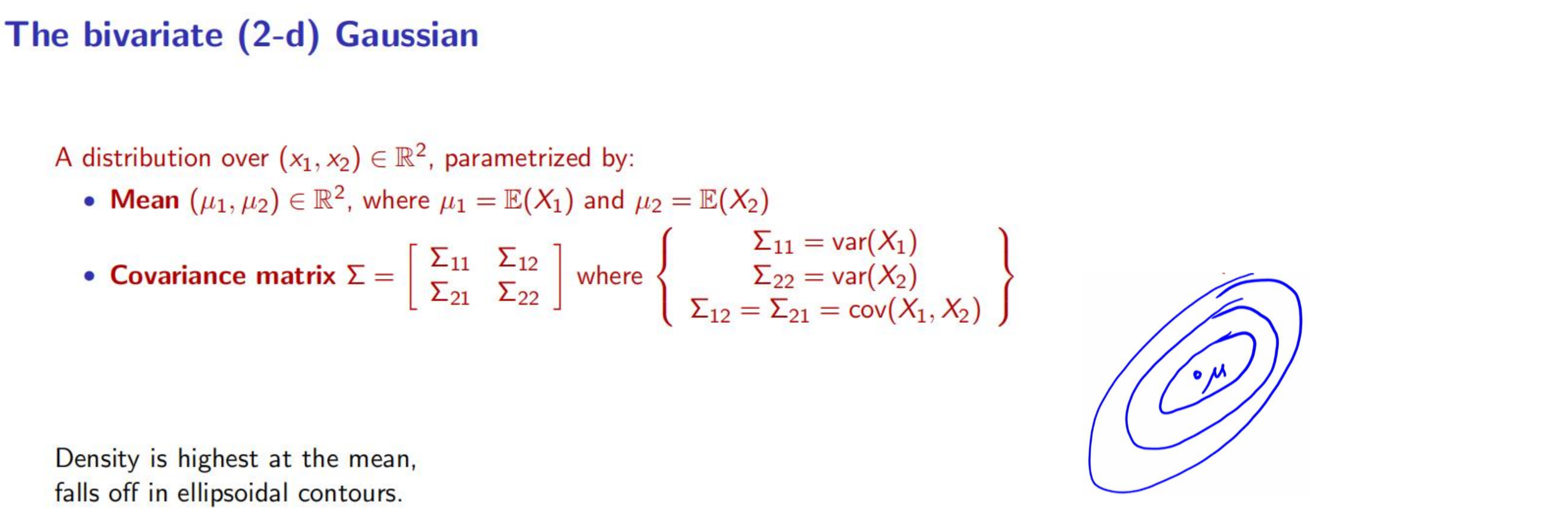
Suppose X_1 has mean μ_1 and X_2 has mean μ_2.

Can measure dependence between them by their **covariance**:

- $\text{cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] = \mathbb{E}[X_1 X_2] - \mu_1 \mu_2$
- Maximized when $X_1 = X_2$, in which case it is $\text{var}(X_1)$.
- It is at most $\text{std}(X_1)\text{std}(X_2)$.

So let's quickly recall what that is. What is the covariance between two random variables? So let's say you have these two random variables, X1 and X2, alcohol and flavonoid, for example. X1 has mean mu1, average alcohol level. X2 has mean mu2, average flavonoid level. A standard measure of the dependence between X1 and X2 is the covariance. It is the difference between the expected value of X1 times X2 and the expected value of X1 times the expected value of X2.

When the covariance is positive, it means there's a positive correlation. When the covariance is negative, it means there's a negative correlation, and if it's zero, it means the two variables are uncorrelated.



The bivariate (2-d) Gaussian

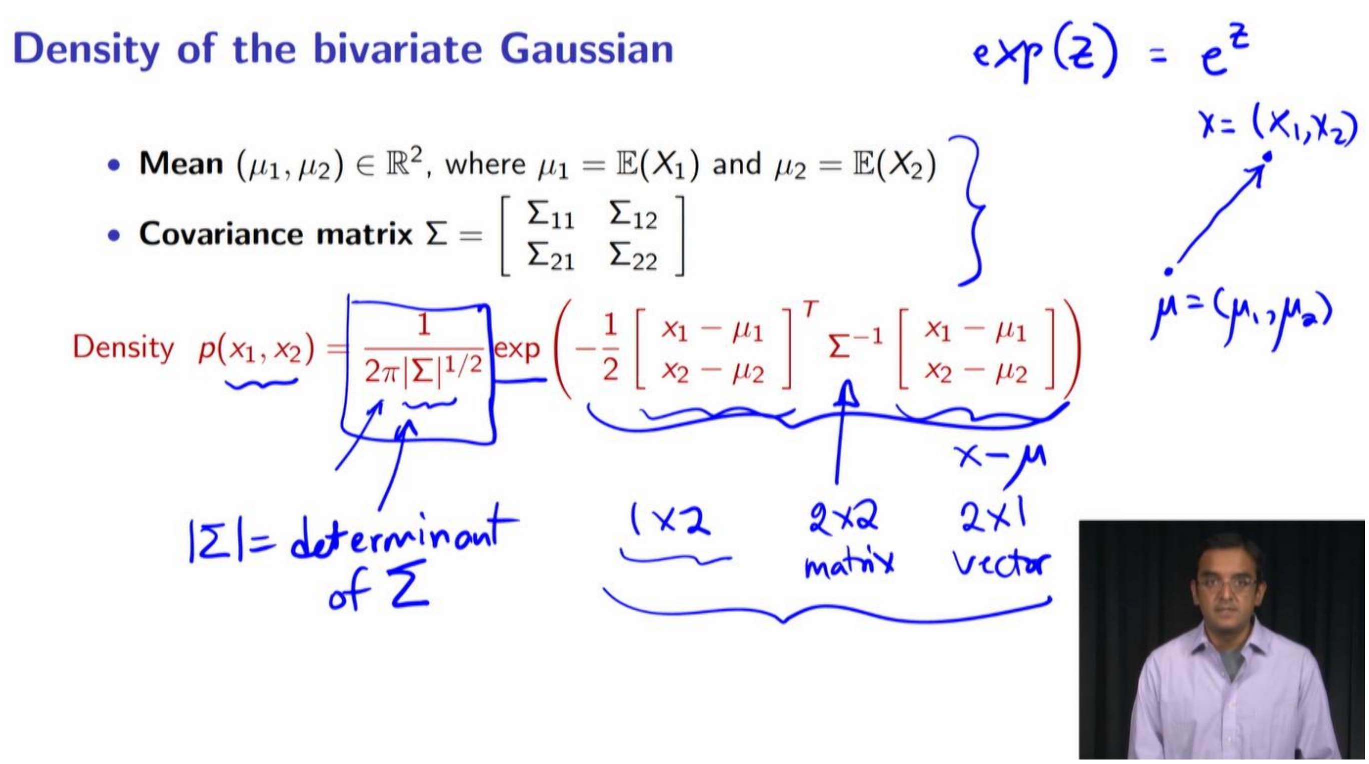
A distribution over $(x_1, x_2) \in \mathbb{R}^2$, parametrized by:

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ where $\left\{ \begin{array}{c} \Sigma_{11} = \text{var}(X_1) \\ \Sigma_{22} = \text{var}(X_2) \\ \Sigma_{12} = \Sigma_{21} = \text{cov}(X_1, X_2) \end{array} \right\}$

Density is highest at the mean,
falls off in ellipsoidal contours.

So with this in mind, let's go and look at the Gaussian distribution a little bit more formally. So the two-dimensional Gaussian is a distribution over 2D space. It's a distribution over all of 2D space, over all points X1, X2. So to every point, it assigns some density. Now the point of highest density is the mean, mu1, mu2, so that's a 2D point. Mu1 is the average value of the X1. Mu2 is the average value of the X2. That's the point of highest density. The other parameters to this distribution are variance and covariance parameters. It's a two-by-two matrix that contains four numbers. But they're actually just three distinct numbers, because it's a symmetric matrix in which the two off-diagonal numbers are identical to each other, okay. So the number on the top left is the variance of X1. The number in the bottom right is the variance of X2. And the number that captures dependence, there's just one number there, that's sigma one-two or sigma two-one, and that's the covariance of X1 and X2. So these are the parameters of the 2D Gaussian distribution.
Now we already talked about the shape of this distribution. The point of highest density is the mean mu. And then the density falls off with these ellipsoidal contours, something like this, where the shape of these ellipsoids is given by the sigma matrix.

总结:



Density of the bivariate Gaussian

$\exp(z) = e^z$

- **Mean** $(\mu_1, \mu_2) \in \mathbb{R}^2$, where $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$
- **Covariance matrix** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

$X = (X_1, X_2)$

$\mu = (\mu_1, \mu_2)$

Density $$p(x_1, x_2) = \boxed{\frac{1}{2\pi|\Sigma|^{1/2}}} \exp\left(-\frac{1}{2} \underbrace{\underbrace{\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T}_{1\times 2} \underbrace{\Sigma^{-1}}_{\substack{2\times 2 \\ \text{matrix}}} \underbrace{\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}}_{\substack{x-\mu \\ 2\times 1 \\ \text{vector}}}}\right)$$

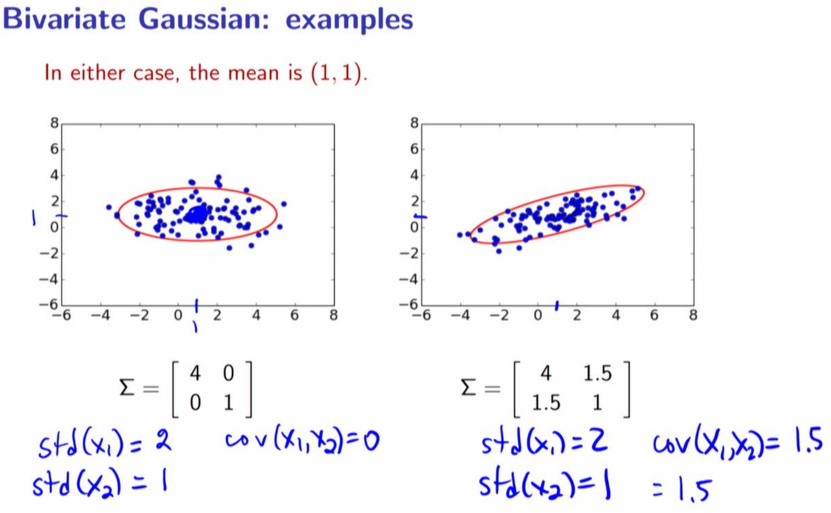
$|\Sigma|$ = determinant of Σ

So now let's look at the precise form of the density. So we've already discussed the parameters. What is the formula for the density? What is the density at any given point X1, X2? Well, it has this formula over here, which admittedly looks a little bit messy. And there might be some terms that are confusing, so let's just go through them one by one.

- So two, I think we can all agree on what that is. Pi, this is just the pi 3.14. It's that pi, okay.
- What is this? So sigma is a two-by-two matrix. What does it mean to put two bars around sigma? Well, that actually is the determinant of the matrix. It's a way to write the determinant of sigma. Some of you might be familiar with this, others not. In either case, don't worry about it. It's something that we'll be coming to again fairly soon. **The important thing to note about this leading term over here is that it does not depend on X at all. It's just a normalizing constant. It's just something that you have to stick in the front to make the density sum up to one**.
- The really important stuff is happening in here, inside the exponent. And by the way, in case you haven't seen this before, the notation, exp(z), is just another way of writing e to the z, okay. So the really important dependence arises from the exponent. And let's look at those terms one by one, okay.
- So in the middle, we have sigma inverse. So sigma is a two-by-two matrix. Its inverse is also a two-by-two matrix. So this is a two-by-two matrix. What's this thing over here, X1 minus mu1, X2 minus mu2, okay? Well, the center of the distribution is some point mu, and we're trying to figure out the density at some other point X. So it turns out that the density depends only on the displacement of X from mu. It depends only on this displacement. And that's what this vector is, over here. It's a two-by-one vector. It's just X minus mu. It's the displacement of X from mu. Mu is the center of the distribution. X is some point whose density we're trying to figure out.
- Now this is also X minus mu, but we've taken its transpose,so it's now a row, so it's one-by-two. And if this seems a little bit confusing, these matrix inverses and transposes and so on, again, you know, don't be concerned. We will be going over these fairly soon. But for the moment, we are multiplying together something that's one-by-two with something that's two-by-two with something that's two-by-one, and so the answer is just one-by-one. It's just a number.

Now the smallest that number can be is zero, and that's what happens when X is identically mu, when you're just measuring the density at the center. And that's when the density is highest. As X moves away from mu, this number gets larger and larger, and the density drops off. So this is a little bit of insight into the density of the bivariate Gaussian.

总结：



Bivariate Gaussian: examples

In either case, the mean is $(1, 1)$.

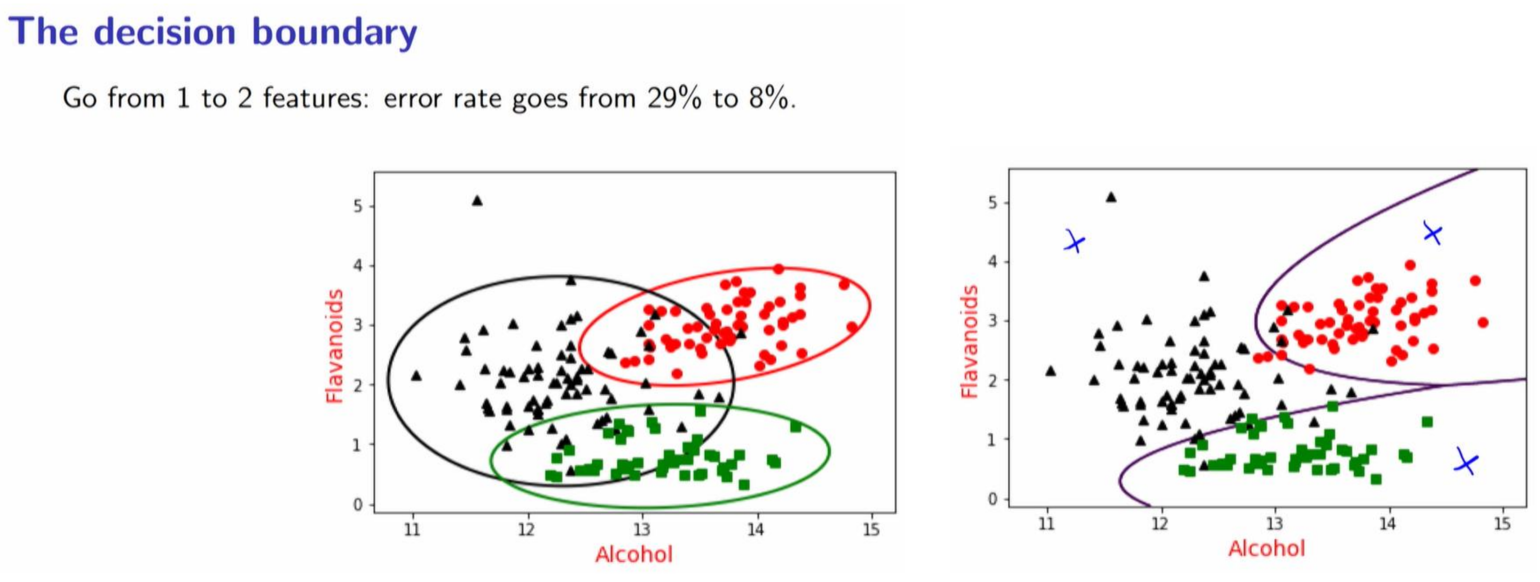
$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 4 & 1.5 \\ 1.5 & 1 \end{bmatrix}$$

$std(x_1) = 2$ $cov(x_1, x_2) = 0$
$std(x_2) = 1$

$std(x_1) = 2$ $cov(X_1, X_2) = 1.5$
$std(x_2) = 1$ $= 1.5$

So let's look at a few examples, just to solidify some of this, okay. So what I've done here is to draw pictures of two different Gaussians with the same mean. So in both cases, the mean is the the point one one, but they have different covariance matrices and therefore different shapes. So for each of them, I picked 50 points at random from the Gaussian and plotted them. Then I've also drawn a representative contour line.

- Okay, so let's look at the one on the left for starters. So the mean is one one. So that's one one, and that is, of course, the point right in the middle over there, okay. The center of the distribution, that's the mean. Okay, so that all makes sense. Now let's look at the numbers in the covariance matrix. Now the number at the upper left is the variance of x one, which is four, so we know the standard deviation of x one is the square root of the variance, so it's two. And so the standard deviation of x two, well, the variance of x two is one, so the standard deviation is one, okay. And the covariance between x one and x two is zero. So what do these tell us? So first of all, the standard deviation of x one is twice that of x two. So the data is twice as spread in the x one direction as it is in the x two direction, and we can see that directly from the plot, okay. The other thing is that the covariance is zero. So x one and x two are uncorrelated. And again, we can see that, because the Gaussian, the data, is not tilting either upwards or downwards. It is axis-aligned, okay.
- So now let's move on to the second picture. So this is a Gaussian with exactly the same mean, one one. The standard deviation of x one is, again, two. The standard deviation of x two is, again, one. So the stretch in the x one direction is, again, twice the stretch in the x two direction. But now, there's a covariance. The covariance between x one and x two is 1.5, which means that there is a bit of a correlation between them, okay. So x one and x two are correlated with each other.



The decision boundary

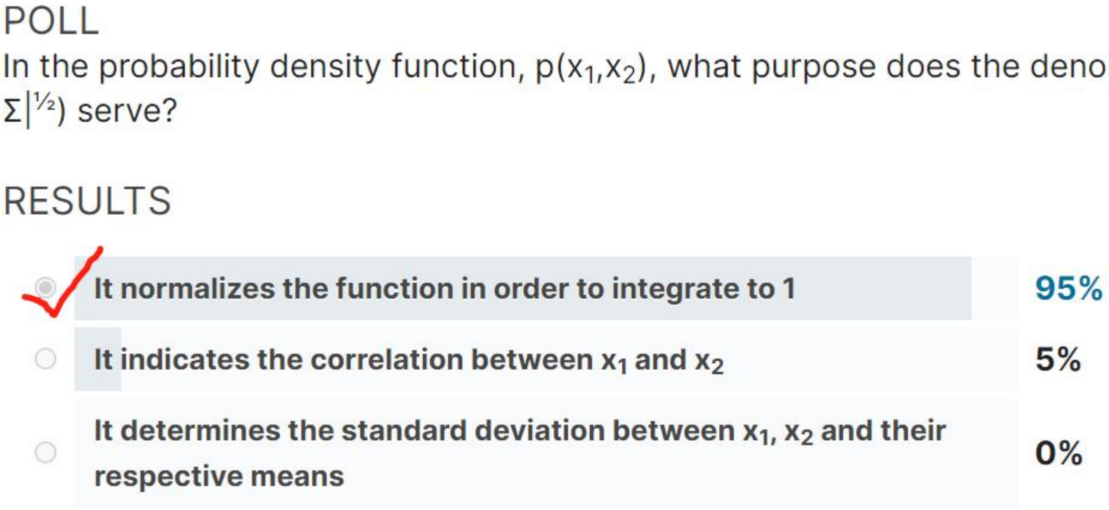
Go from 1 to 2 features: error rate goes from 29% to 8%.

What kind of function is this? And, can we use more features?

So what does that decision boundary look like? It turns out this is what it is, okay. Any point in here gets predicted as being from winery number one. Any point in here gets predicted as being from winery number three, the green one. And any point out here gets predicted as being from winery number two. It's rather a strange decision boundary, isn't it?

What is the functional form of this boundary? Now that we've reduced the error to 8%, by using two features, can we do better by using more features? How do we do that? How do we take care of correlations between multiple variables? So we'll see the answers to all these questions very soon. See you next time.

总结：



POLL

In the probability density function, $p(x_1,x_2)$, what purpose does the denominator $1/(2\pi|\Sigma|^{1/2})$ serve?

RESULTS

● ✓	It normalizes the function in order to integrate to 1	95%
○	It indicates the correlation between x_1 and x_2	5%
○	It determines the standard deviation between x_1, x_2 and their respective means	0%

总结：