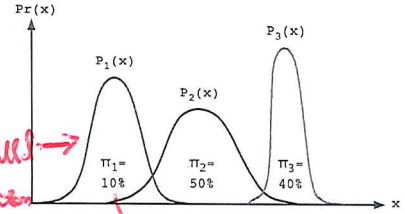
WEEK 2-1 14/June/2023.

The generative approach to classification

- Generative approach based on probability distributions.
- Main idea of Generative approach:
  fit each class seperately with a probability distribution.

Generative models

Example:
- Data space $\mathcal{X} = \mathbb{R}$
- Classes/labels $\mathcal{Y} = \{1, 2, 3\}$
- first look at the points labelled '1' and we fit a probability distribution to them.



- more labeled '1' makes up 10% of the training set. $\pi_1 = 1/10$.

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

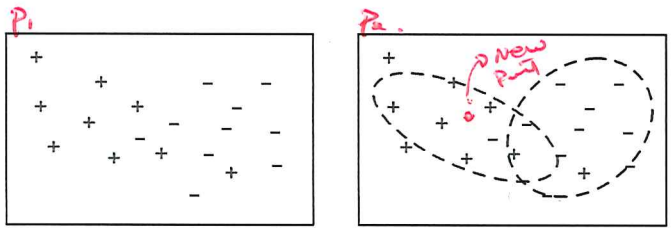
Overall **joint distribution**: $\Pr(x, y) = \Pr(y)\Pr(x|y) = \pi_y P_y(x)$.

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

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

if it is the largest than label y for new x is 3.

The generative approach to classification



P1: Training Set with 15-20 points
P2: 2 labels. plus / minus
P3: look at pluses label and we fit a model to them
P4: then look at minus we ...
P5: Ellipse shaped distribution that fit into pluses & minuses.

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?

WEEK 2-2 14/June/2018.

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**.

Probability spaces

You roll two dice.

Q. What is the probability they add to 10?

The **probability space** has two components:

1. **Sample space** (space of outcomes). the set of all possible outcomes.

$\Omega = \{(1,1),(1,2)\dots(1,6),(2,1),\dots,(6,6)\}$

$= \{1,2,\dots 6\} \times \{1,2\dots 6\} = \{1,2,3,4,5,6\}^2 = 36.$

2. **Probabilities of outcomes**, summing to 1. Probability of each of these outcomes.

- the number of possible outcomes = $6\times6 = 36$.
- Each outcome has prob $1/36$.

↳ probability spaces.

Events

Probability space:

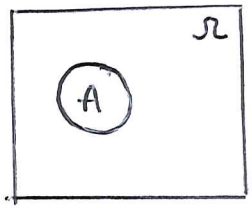
- Outcomes: $\Omega = \{\text{all possible pairs of dice rolls}\}$
- Every pair $z = (z_1, z_2) \in \Omega$ has probability 1/36.

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

$A \subseteq \Omega$ (An Event is just a subset of the possible outcomes).

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

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



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

3 outcomes; each of them has probability 1/36.

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?

sample space. $\Omega = \{1, 2, \ldots 10\} \times \{H, T\} \times \{H, T\} \times \{H, T\} \times \{H, T\}$. 10 coins, 4 tosses.

$= \{1, 2, \ldots 10\} \times \{H, T\}^4 = 10 \times 2 \times 2 \times 2 \times 2 = 160$

- 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 = picked the bad coin

$= \{(10, -, -, -)\}$

B = all coins are tails

$= \{(-, T, T, T, T)\}$

$\Pr(A \cap B) = \Pr(\text{bad coin}) \Pr(\text{all tails} \mid \text{bad coin})$

$= 1/10 \times 1$

$= 1/10.$

Conditioning

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

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

↳ "given"

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 $\Pr(A|B)$

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

- 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

$\Pr(\text{all tails}) = \Pr(\text{bad coin} \mid \text{all tails})$

$\Pr(B)$. $\quad + \Pr(\text{not bad coin} \mid \text{all tails})$

$= \frac{1}{10} \times 1 + \frac{9}{10} \times (\frac{1}{2})^4 = \frac{5}{32}$

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

$= \frac{1/10}{5/32} = 0.64$

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?

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

correction factor

$$\Pr(A|B) = \frac{\Pr(A \wedge B)}{P(B)} = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$$

Prob. If a coin is flipped 5 times, what is the size of the sample space for the experiment?

- The sample space of a fair coin flip is $\{H, T\}$.
- The sample space of a coin flips is all 2^5 possible sequences of outcomes $2^5 = 32$.

WEEK-3. 15/June/2018.

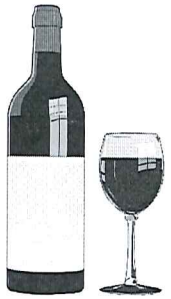
Generative modeling in one dimension

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.

The data set

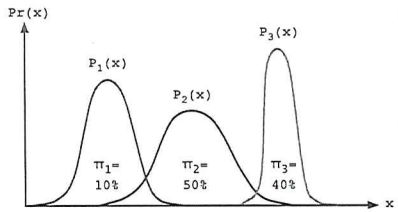
Training set obtained from 130 bottles Train Set.

- Winery 1: 43 bottles
- Winery 2: 51 bottles
- Winery 3: 36 bottles
- For each bottle, 13 features: 13 dimensional vectors.
  '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. Test Set.

- use this data to build classifier that takes the features from a new bottle and predict its label.

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 $\pi_j P_j(x)$.

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'

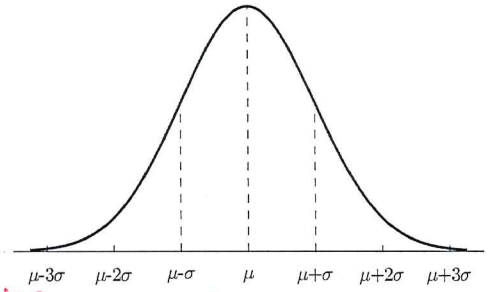
Class weights:

$\pi_1 = 43/130 = 0.33$, $\pi_2 = 51/130 = 0.39$, $\pi_3 = 36/130 = 0.28$

Need distributions P_1, P_2, P_3, one per class.

Base these on a single feature: 'Alcohol'.

The univariate Gaussian

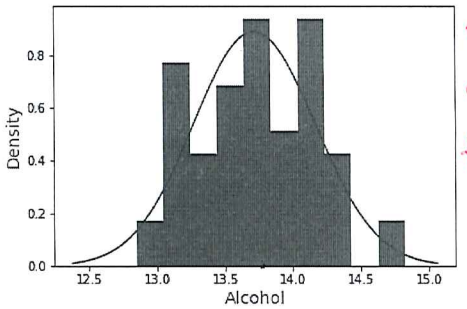


The Gaussian $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

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

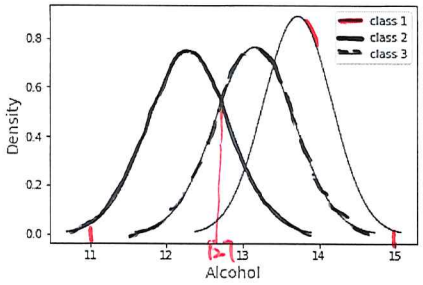
The distribution for winery 1

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 $\pi_j P_j(x)$

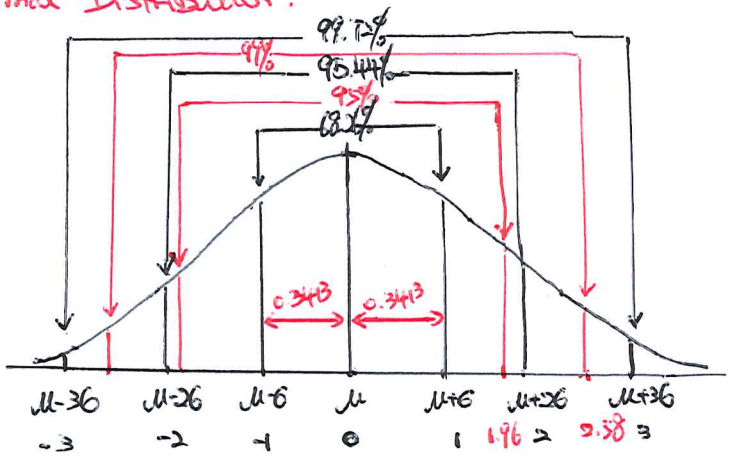
Test error: 14/48 = 29%

$x = 15$ label 1

$x = 11$ label 2

$x = 12.7$ label 3 ($\pi_2 > \pi_3$)

Normal Distribution.



- $z = \frac{x - \mu}{\sigma}$

$x_1 = 1 \quad x_2 = -1$

$z_1 = \frac{1-0}{1} = 1 \quad z_2 = \frac{-1-0}{1} = -1$ look for the z Table find 0.3413 (34.13%)

$34.13\% \times 2 = 68.26\%$

The percentage of people that have x between 1 and -1 is 68.26% of the populations.

- $\mu = 0$. mean = median $\sigma = 1$
- $\mu + \sigma = 0 + 1 \quad \mu - \sigma = -1$
- 68.26% - in total, observations within plus or minus one stddev away from the mean (-1/3) ; ~~most people are between~~
- 1.96 - significant score, corresponds to 95% of the total distribution; so there are 5% left on the 2 tails (0.05).
- 2.58 - corresponds to 99% of total distribution.

week 2 - 4 15/June/2018.

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

Random variables

Roll two dice. Let X be their sum.

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

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 information in the underlying probability space.
Q. How we can determine the distribution of a random variable from

The distribution of a random variable

Roll a die.

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

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

$\Pr(X=0) = \Pr(\text{die} = 1, 2) = \frac{2}{6} = \frac{1}{3}$

$\Pr(X=1) = \Pr(\text{die} = 3, 4, 5, 6) = \frac{4}{6} = \frac{2}{3}$

Expected value, or mean

Expected value of a random variable X:

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

The expected x is simply the value of x if we repeated the experiment over & over again.

It is a weighted average of all possible values that x can take.

where each value is weighted by its probability of occurring.

Roll a die. Let X be the number observed.
What is $\mathbb{E}(X)$?

$x \in \{1, 2, 3, 4, 5, 6\}$

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

$= 3.5$

Another example

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

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

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

$= p.$

A property of expected values

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

- You double the numbers? - The average also get doubled.
- You increase each number by 1? - the average also increase by 1.

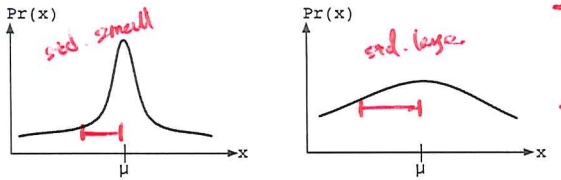
Summary: Let X be any random variable.
If $V = aX + b$ (any constants a, b), then $\mathbb{E}(V) = a\mathbb{E}(X) + b$

- a new random variable V.
- V to be some constant time X plus some other constant.

↳ Linearity property

Variance

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



- Two distributions with same mean μ.
- Left, tightly concentrated around its mean; Right, more diffused.

Expected value of abs distance from X to μ.

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

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

- How can we capture the spread? Using the average distance from the mean.

Variance: example

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

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

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

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

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

Alternative formula for variance

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

Another way to write it: $\text{var}(X) = \mathbb{E}(X^2) - \mu^2$ → minus mean squared.

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

↓ expected value of x squared.

$\mathbb{E}(X^2) = 1^2 \cdot \frac{1}{5} + 2^2 \cdot \frac{1}{5} + 3^2 \cdot \frac{1}{5} + 4^2 \cdot \frac{1}{5} + 5^2 \cdot \frac{1}{5} = 11$

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

Variance: properties

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

- Variance is always ≥ 0 non-negative number.
- How is the variance affected if:
  - You increase each number by 1? variance is unaffected (everything shifts over a little bit)
  - You double each number? std is double, variance is $\times 4$ (but the spread remains the same).
- Summary: If $V = aX + b$ then $\text{var}(V) = a^2 \text{var}(X)$

Week-2-S 18/Jun/20 S.

Probability review III: Measuring dependence

- when dealing with multiple random variables.
  eg., a patient's heart rate, blood pressure... the easiest situation is all variables are independent.
- Then don't worry about the interactions between them.

Topics we'll cover

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

Independent random variables

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

- X being equal to little x doesn't either make it more likely or less likely that Y being equals to little y.

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

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

$\Pr(X = \heartsuit) = \frac{1}{4}$

$\Pr(Y = 6) = \frac{1}{13}$

$\Pr(X = \heartsuit, Y = 6) = \frac{1}{4} \times \frac{1}{13} = \frac{1}{52}$

Q. Are X and Y indep?

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.

↓
the total number of heads.

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

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

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

∴ Not independent variables.

Q. Are X and Y indep?

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

X	Pr
-1	0.8
0	0.1
1	0.1

Y	Pr
-1	0.5
0	0.2
1	0.3

← To check if these are independent, we have to check every entry in this table of joint distribution.

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

$= 0.8 \times 0.3$

∴ independent X and Y.

Dependence

Example: Pick a person at random, and take

$H =$ height
$W =$ weight

Independence would mean

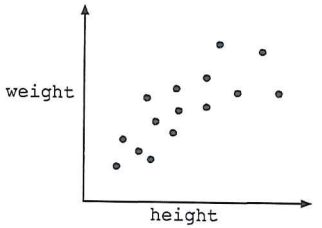
$$\Pr(H = h, W = w) = \Pr(H = h)\Pr(W = w).$$

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

- we would expect height and weight to be positively correlated.

Positive correlation

H, W are **positively correlated**



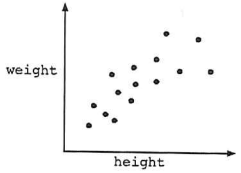
- A general upward trend. People who are taller tend to be heavier.

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

the average of H times W.

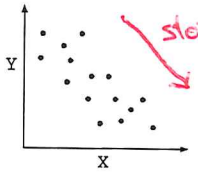
Average H times Average W.

Types of correlation



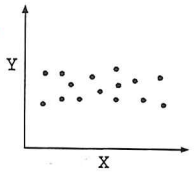
H, W **positively correlated**
This also implies

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



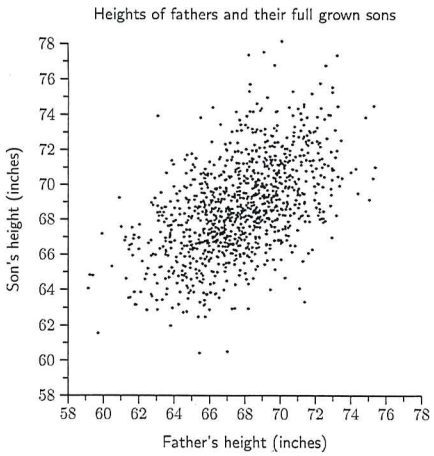
sloping downward trend.

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



- Carl Pearson by 20th century
- general upward trend, seem to be a positive correlation
- Correlation coefficient

$r \in [-1, 1]$

↓

Perfect negative correlation.

Covariance and correlation

- Covariance

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

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

- Correlation

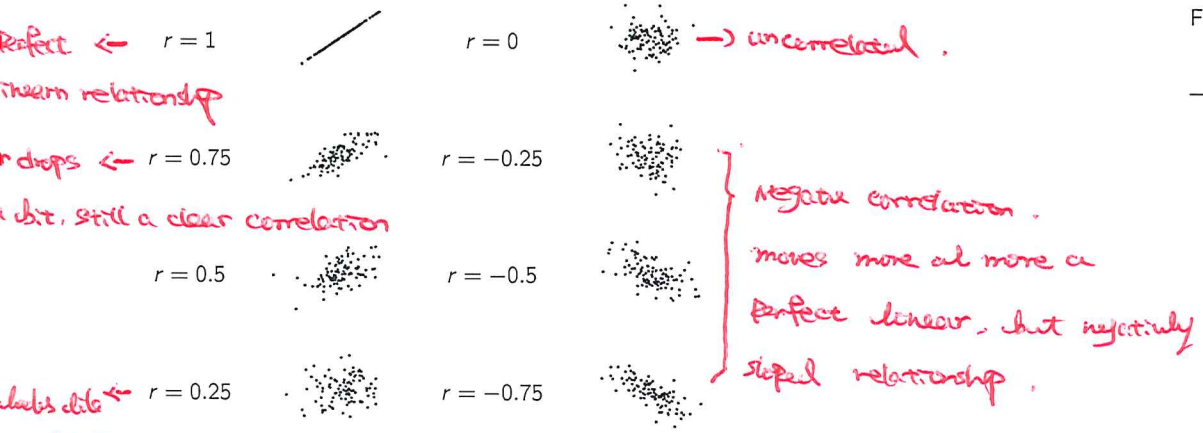
$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\text{std}(X)\text{std}(Y)}$$

Pearsons correlation coefficient.

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

If X, Y independent then $\text{cov}(X, Y) = 0$. — when 2 variables are independent, they are gonna be uncorrelated,
But the converse need not be true. — Uncorrelated doesn't mean independent. independent means uncorrelated.

Correlation coefficient: pictures



Q. what's the formular? obtain the value r?

Covariance and correlation: example

Find $\text{cov}(X, Y)$ and $\text{corr}(X, Y)$

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

$\text{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$ $\quad \text{Cor}(X,Y) = \frac{\text{Cov}(X,Y)}{\text{Std}(X)\text{Std}(Y)}$

$= -1 - 0 = -1$ $\quad = -\frac{1}{3}$

XY	Pr
−3	2/3
3	1/3

$\mathbb{E}(XY) = -\frac{2}{3} + \frac{1}{3} = -1$

∴ weak negatively correlations.

X	Pr
−1	1/2
1	1/2

$\mathbb{E}(X) = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0$

$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 1 - 0 = 1$

$\text{Std}(X) = 1$. Distance of X from mean is always 1.

Y	Pr
−3	1/2
3	1/2

$\mathbb{E}(Y) = 0$

$\text{Var}(Y) = 9 - 0 = 9$.

$\text{Std}(Y) = 3$. typical distance from Y to its mean is 3.

week 2-6 19 June 2018

Two-dimensional generative modeling with the bivariate Gaussian

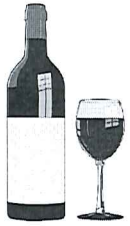
- look at Gaussian in 2 dimensions. allows us to model the dependence between features.

Topics we'll cover

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

The winery prediction problem

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



one feature used

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

What if we use **two** features? - error rate drops.

The data set, again

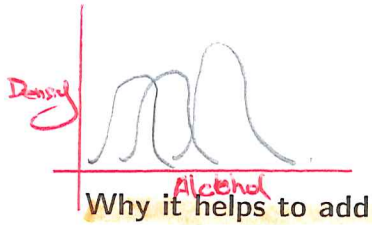
Training set obtained from 130 bottles

- 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'.

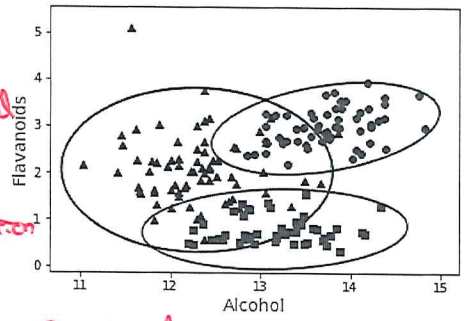
Q. why it might be helpful to throw in a 2nd feature?



- Fit a Gaussian to alcohol level from class 1;
" " " class 2;
" " " class 3.
- the 3 Gaussians aren't separated very well: one feature alone is not a great basis for doing classification.

Why it helps to add features

Better **separation** between the classes!



- for each point, they have 2 features, hence in 2 dimensions.
- Quite a bit of separation between them
- Two features may be better for classification.

Error rate drops from 29% to 8%.

- Error rate drops off to 8%.

Dependence between two random variables

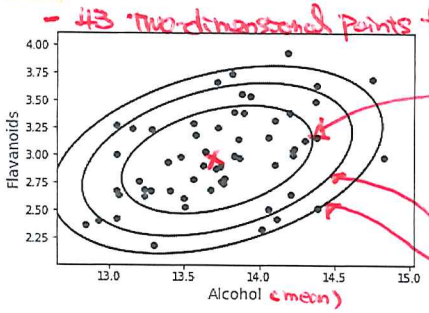
Suppose X_1 has mean μ_1 and X_2 has mean μ_2.

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 $\text{std}(X_1)\text{std}(X_2)$.

- Cor = positive ~ positive correlation
- Cor = negative ~ negative correlation
- Cor = 0 ~ non correlation

The bivariate Gaussian



- 43 two-dimensional points from class 1.
- let's say density is 0.5 at mean.
this ellipse, the points in this density 0.4.
density 0.3.
density 0.2.
- Fit a Gaussian distribution with 2 parameters.

Model class 1 by a bivariate Gaussian, parametrized by:

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

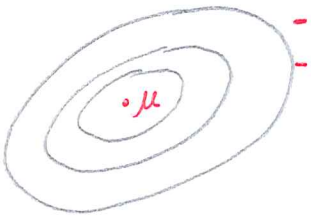
- Covariance between 2 features.
- variance along the 2nd direction
- variance along x one direction
- Center of distribution.
- Highest density.
- As move away from mean, the density drops off, and it drops off with the ellipsoidal contours.

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.



- μ: the highest density
- the shape of ellipsoids is given by sigma matrix

Density of the bivariate Gaussian

- **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** $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

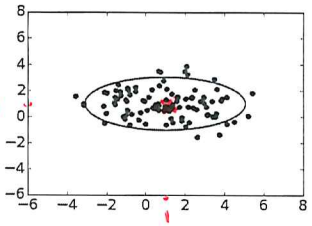
} Parameters.

Density $p(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{bmatrix}\right)$

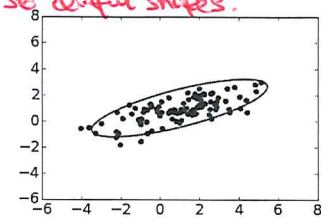
- $x = (x_1, x_2)$
- $\mu (\mu_1, \mu_2)$
- displacement of x from μ.
- A 2x1 vector
- $x - \mu$
- A 2x2 matrix
- A 1x2 matrix
- $|\Sigma|$ = determinant of Σ
- exp(x) = e^x
- 3.14
- Not dependent on x; just a normalizing constant.
- Just a number
- can be 0, when x is identically μ. when measuring the density at center. the highest density
- move away from us density ~~is larger~~ drops off, number is getting larger.

Bivariate Gaussian: examples

In either case, the mean is (1, 1). Diff but Covariance matrices: so diff shapes.



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



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

std(X1) = $\sqrt{4}$ = 2

std(X2) = $\sqrt{1}$ = 1

corr(X1, X2) = 0 - uncorrelated, neither upwards nor downwards.

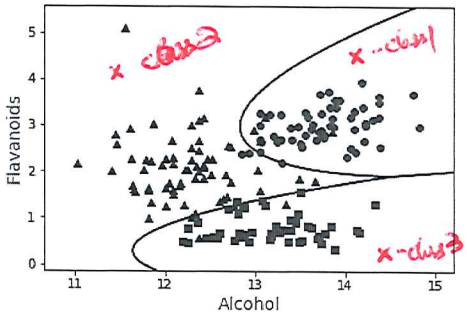
std(X1) = 2 std(X2) = 1 ⇒ the stretch in X one direction is twice the stretch in the X two direction.

corr(X1, X2) = 1.5 ⇒ correlated.

- The data is twice a spread in the X one direction as it is in the X two direction.

The decision boundary

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



- when a new point comes.
- Bayes's Rule.
- $\pi_y P_y(x)$ maximized.

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