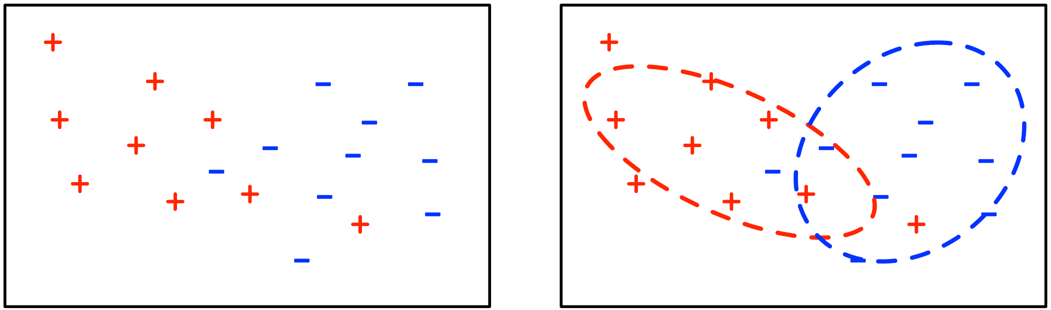
The generative approach to classification

The generative approach to classification



The learning process:

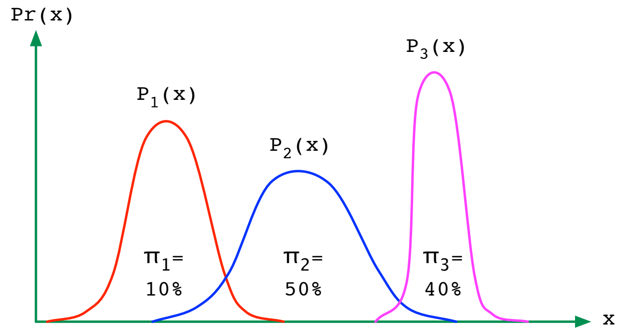
- Fit a probability distribution to each class, individually

To classify a new point:

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

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 $\Pr(x, y)$

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.

What is the probability they add to 10?

The **probability space** has two components:

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

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

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.

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?

- 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.

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 $\Pr(A|B)$

- 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

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?

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

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.

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

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 $X = 0$.

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 $\mathbb{E}(X)$?

Another example

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

A property of expected values

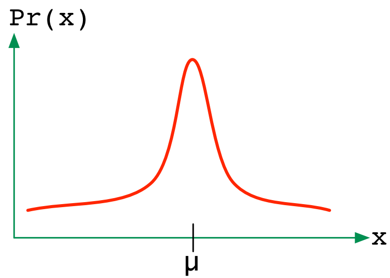
How is the average of a set of numbers affected if:

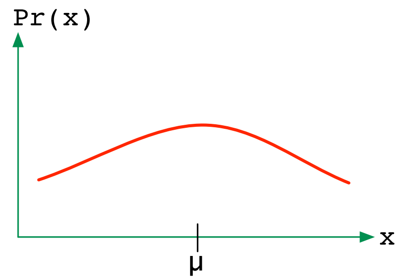
- You double the numbers?
- You increase each number by 1?

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

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.

Variance: example

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

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 $\mathrm{var}(V) = a^2\,\mathrm{var}(X)$

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 $\{1, 2, 3, 4, 5\}$.

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**

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.

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.

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

Dependence

Example: Pick a person at random, and take

$$H = \text{height}$$
$$W = \text{weight}$$

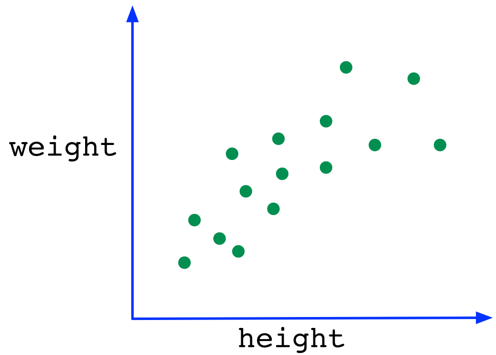
Independence would mean

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

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

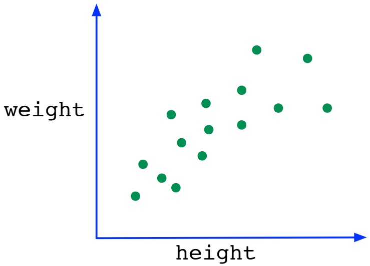
Positive correlation

H, W are **positively correlated**



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

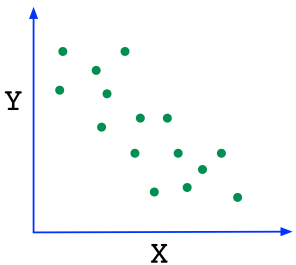
Types of correlation



H, W **positively correlated**

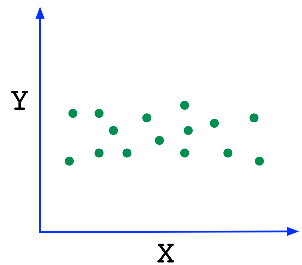
This also implies

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



X, Y **negatively correlated**

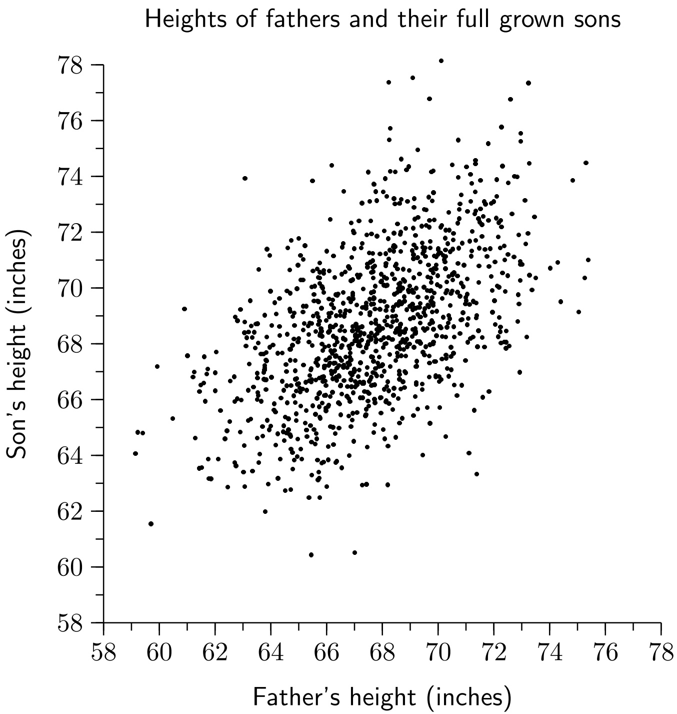
$\mathbb{E}[XY] < \mathbb{E}[X]\,\mathbb{E}[Y]$



X, Y **uncorrelated**

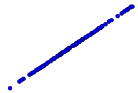
$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$

Pearson (1903): fathers and sons

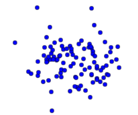


Correlation coefficient: pictures

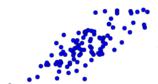
$r = 1$



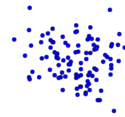
$r = 0$



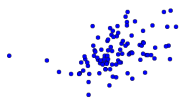
$r = 0.75$



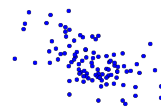
$r = -0.25$



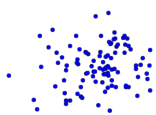
$r = 0.5$



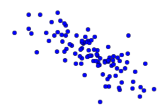
$r = -0.5$



$r = 0.25$



$r = -0.75$



Covariance and correlation

- Covariance

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

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

- Correlation

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

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

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

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	$1/6$

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.

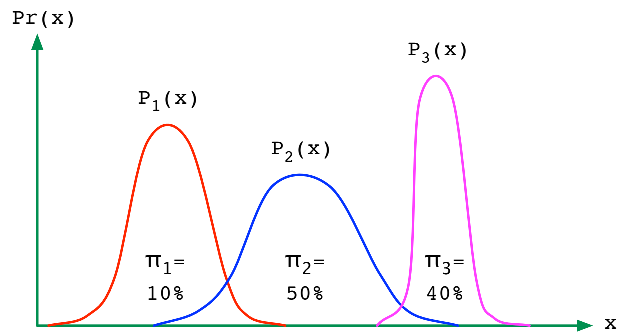
The data set

Training set obtained from 130 bottles

- 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 $\pi_j P_j(x)$.

Fitting a generative model

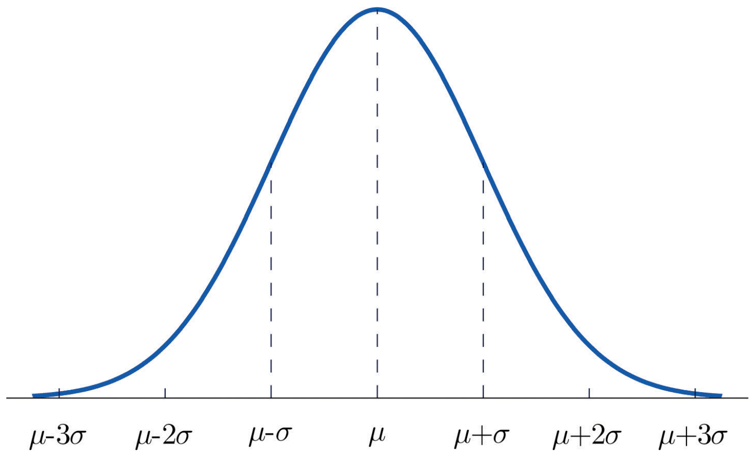
Training set of 130 bottles:

- 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'.

The univariate Gaussian

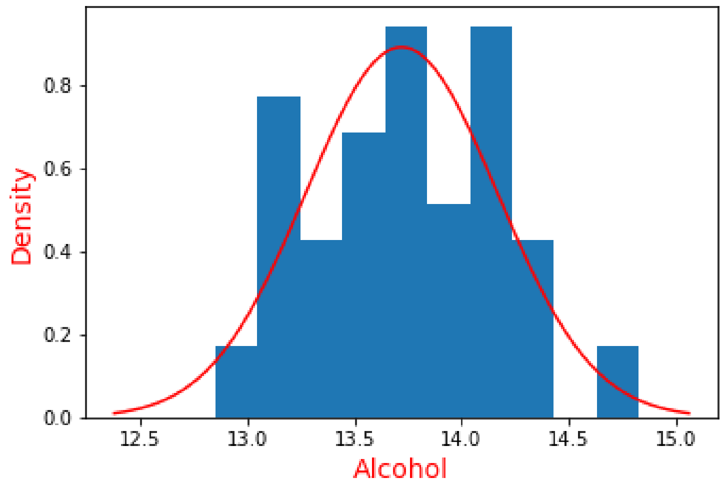


The Gaussian $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

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

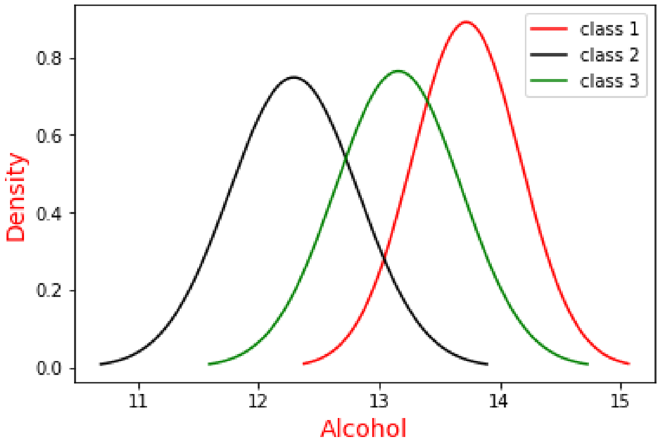
The distribution for winery 1

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: $14/48 = 29\%$

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

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

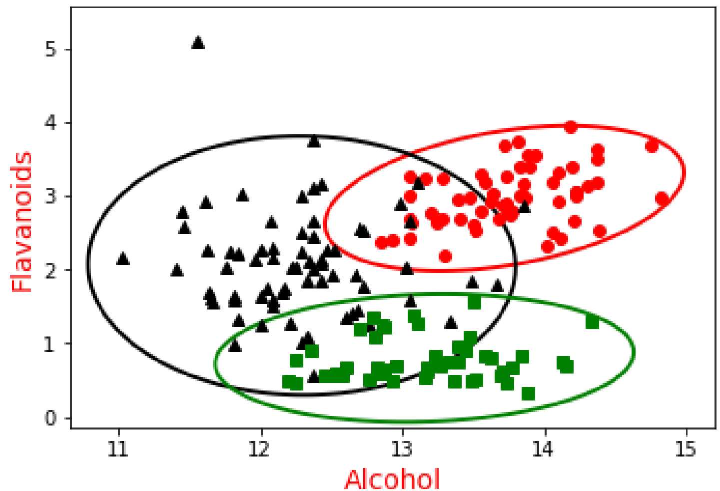
- Winery 1: 43 bottles
- Winery 2: 51 bottles
- Winery 3: 36 bottles
- For each bottle, 13 features:
  'Alcohol', 'Malic acid', 'Ash', 'Alcalinity of ash','Magnesium',
  'Total phenols', 'Flavanoids', 'Nonflavanoid phenols', 'Proanthocyanins',
  'Color intensity', 'Hue', 'OD280/OD315 of diluted wines',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

Also, a separate test set of 48 labeled points.

This time: 'Alcohol' and 'Flavanoids'.

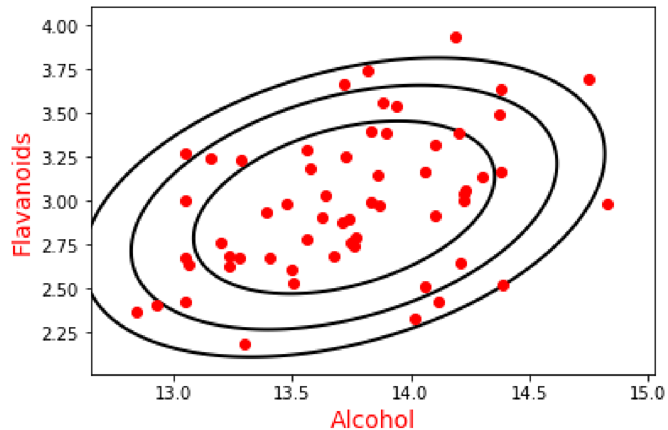
Why it helps to add features

Better **separation** between the classes!



Error rate drops from 29% to 8%.

The bivariate Gaussian



Model class 1 by a bivariate Gaussian, parametrized by:

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

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

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** $\Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array}\right]$ 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.

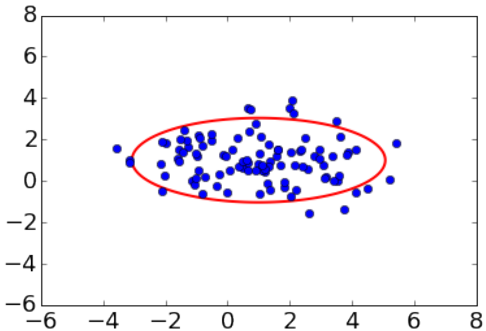
Density of the bivariate Gaussian

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

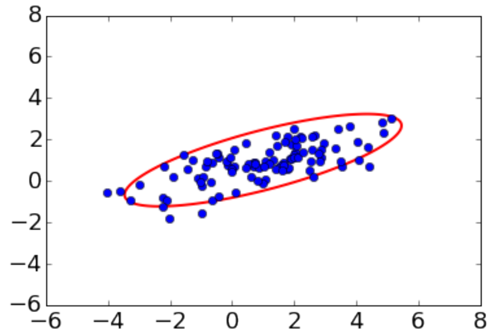
Density $p(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \Sigma^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}\right)$

Bivariate Gaussian: examples

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



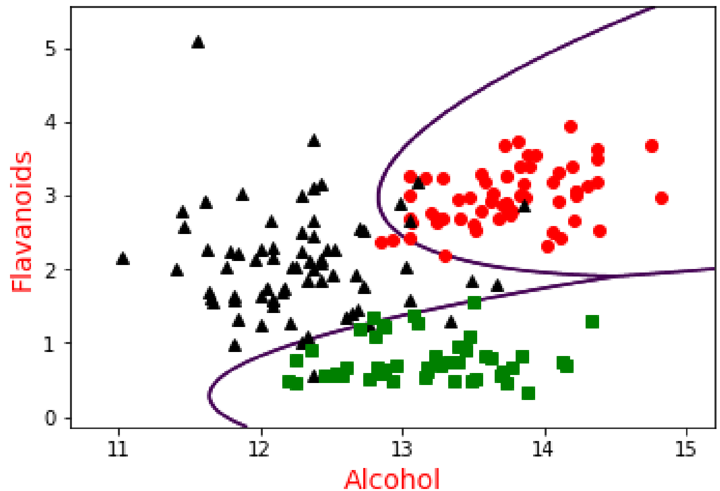
$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}$$



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

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?