Random variables, expectation, and variance

DSE 210

Random variables

Roll a die.

$$\text{Define } X = \begin{cases} 1 & \text{if die is} \geq 3 \\ 0 & \text{otherwise} \end{cases}$$

Here the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.

$$\omega = 1, 2 \Rightarrow X = 0$$
$$\omega = 3, 4, 5, 6 \Rightarrow X = 1$$

Roll n dice.

$$X = \#\text{ of 6's}$$
$$Y = \#\text{ of 1's before the first 6}$$

Both X and Y are defined on the same sample space, $\Omega = \{1, 2, 3, 4, 5, 6\}^n$. For instance,

$$\omega = (1, 1, 1, \ldots, 1, 6) \;\Rightarrow\; X = 1, Y = n - 1.$$

In general, a **random variable (r.v.)** is a defined on a probability space. It is a mapping from Ω to $\mathbb{R}$. 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 takes values in $\{0, 1\}$ and has distribution:

$$\Pr(X = 0) = \frac{1}{3} \text{ and } \Pr(X = 1) = \frac{2}{3}.$$

Roll n dice. Define $X =$ number of 6's.

X takes values in $\{0, 1, 2, \ldots, n\}$. The distribution of X is:

$$\begin{aligned}\Pr(X = k) &= \#(\text{sequences with } k \text{ 6's}) \cdot \Pr(\text{one such sequence}) \\ &= \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}\end{aligned}$$

Throw a dart at a dartboard of radius 1. Let X be the distance to the center of the board.

X takes values in $[0, 1]$. The distribution of X is:

$$\Pr(X \leq x) = x^2.$$

Expected value, or mean

The expected value of a random variable X is

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

Roll a die. Let X be the number observed.

$$\begin{aligned}\mathbb{E}(X) &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} \\ &= \frac{1+2+3+4+5+6}{6} = 3.5 \quad \text{(average)}\end{aligned}$$

Biased coin. A coin has heads probability p. Let X be 1 if heads, 0 if tails.

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Toss a coin with bias p repeatedly, until it comes up heads.
Let X be the number of tosses.

$$\mathbb{E}(X) = \frac{1}{p}.$$

Pascal's wager

Pascal: I think there is some chance ($p > 0$) that God exists. Therefore I should act as if he exists.

Let X = my level of suffering.

- Suppose I behave as if God exists (that is, I behave myself).
  Then X is some significant but finite amount, like 100 or 1000.
- Suppose I behave as if God doesn't exists (I do whatever I want to).
  If indeed God doesn't exist: $X = 0$.
  But if God exists: $X = \infty$ (hell).
  Therefore, $\mathbb{E}(X) = 0 \cdot (1 - p) + \infty \cdot p = \infty$.

The first option is much better!

Linearity of expectation

- If you double a set of numbers, how is the average affected?
  It is also doubled.
- If you increase a set of numbers by 1, how much does the average change?
  It also increases by 1.
- Rule: $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ for any random variable X and any constants a, b.
- But here's a more surprising (and very powerful) property:
  $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ for any two random variables X, Y.
- Likewise: $\mathbb{E}(X + Y + Z) = \mathbb{E}(X) + \mathbb{E}(Y) + \mathbb{E}(Z)$, etc.

Linearity: examples

Roll 2 dice and let Z denote the sum. What is $\mathbb{E}(Z)$?

Method 1
Distribution of Z:

z	2	3	4	5	6	7	8	9	10	11	12
$\Pr(Z = z)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Now use formula for expected value:

$$\mathbb{E}(Z) = 2 \cdot \frac{1}{36} + 3 \cdot \frac{2}{36} + 4 \cdot \frac{3}{36} + \cdots = 7.$$

Method 2
Let X_1 be the first die and X_2 the second die. Each of them is a single die and thus (as we saw earlier) has expected value 3.5. Since $Z = X_1 + X_2$,

$$\mathbb{E}(Z) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 3.5 + 3.5 = 7.$$

Toss n coins of bias p, and let X be the number of heads. What is $\mathbb{E}(X)$?

Let the individual coins be $X_1, \ldots, X_n$.
Each has value 0 or 1 and has expected value p.

Since $X = X_1 + X_2 + \cdots + X_n$,

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = np.$$

Roll a die n times, and let X be the number of 6's. What is $\mathbb{E}(X)$?

Let X_1 be 1 if the first roll is a 6, and 0 otherwise.

$$\mathbb{E}(X_1) = \frac{1}{6}.$$

Likewise, define $X_2, X_3, \ldots, X_n$.

Since $X = X_1 + \cdots + X_n$, we have

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n) = \frac{n}{6}.$$

Coupon collector, again

Each cereal box has one of k action figures. What is the expected number of boxes you need to buy in order to collect all the figures?

Suppose you've already collected $i-1$ of the figures. Let X_i be the time to collect the next one.

Each box you buy will contain a new figure with probability $(k-(i-1))/k$. Therefore,

$$\mathbb{E}(X_i) = \frac{k}{k-i+1}.$$

Total number of boxes bought is $X = X_1 + X_2 + \cdots + X_k$, so

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_k) \\ &= \frac{k}{k} + \frac{k}{k-1} + \frac{k}{k-2} + \cdots + \frac{k}{1} \\ &= k\left(1 + \frac{1}{2} + \cdots + \frac{1}{k}\right) \quad \approx \quad k \ln k.\end{aligned}$$

Independent random variables

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

Independent or not?

- Pick a card out of a standard deck. $X =$ suit and $Y =$ number.
  Independent.
- Flip a fair coin n times. $X = \#$ heads and $Y =$ last toss.
  Not independent.
- X, Y take values $\{-1, 0, 1\}$, with the following 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

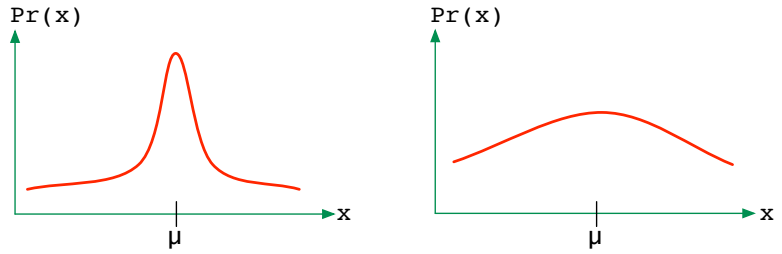
	X	Y
-1	0.8	0.5
0	0.1	0.2
1	0.1	0.3

Independent.

Variance

If you had to summarize the entire distribution of a r.v. X by a single number, you would use the mean (or median). Call it μ.

But these don't capture the *spread* of X:



What would be a good measure of spread? How about the average distance away from the mean: $\mathbb{E}(|X - \mu|)$?

For convenience, take the square instead of the absolute value.

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

where $\mu = \mathbb{E}(X)$. The variance is always ≥ 0.

Variance: example

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

Toss a coin of bias p. Let $X \in \{0, 1\}$ be the outcome.

$$\begin{aligned}
\mathbb{E}(X) &= p \\
\mathbb{E}(X^2) &= p \\
\mathbb{E}(X - \mu)^2 &= p^2 \cdot (1 - p) + (1 - p)^2 \cdot p = p(1 - p) \\
\mathbb{E}(X^2) - \mu^2 &= p - p^2 = p(1 - p)
\end{aligned}$$

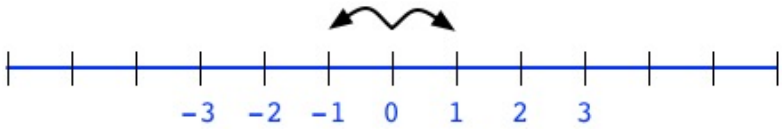
This variance is highest when $p = 1/2$ (fair coin).

The standard deviation of X is $\sqrt{\text{var}(X)}$.
It is the average amount by which X differs from its mean.

Variance of a sum

$\text{var}(X_1 + \cdots + X_k) = \text{var}(X_1) + \cdots + \text{var}(X_k)$ if the X_i are independent.

Symmetric random walk. A drunken man sets out from a bar. At each time step, he either moves one step to the right or one step to the left, with equal probabilities. Roughly where is he after n steps?



Let $X_i \in \{-1, 1\}$ be his ith step. Then $\mathbb{E}(X_i) = ?0$ and $\text{var}(X_i) = ?1$.

His position after n steps is $X = X_1 + \cdots + X_n$.

$$\mathbb{E}(X) = 0$$
$$\text{var}(X) = n$$
$$\text{stddev}(X) = \sqrt{n}$$

He is likely to be pretty close to where he started!

Sampling

Useful variance rules:

- $\text{var}(X_1 + \cdots + X_k) = \text{var}(X_1) + \cdots + \text{var}(X_k)$ if X_i's independent.
- $\text{var}(aX + b) = a^2\text{var}(X)$.

What fraction of San Diegans like sushi? Call it p.

Pick n people at random and ask them. Each answers 1 (likes) or 0 (doesn't like). Call these values $X_1, \ldots, X_n$. Your estimate is then:

$$Y = \frac{X_1 + \cdots + X_n}{n}.$$

How accurate is this estimate?
Each X_i has mean p and variance $p(1-p)$, so

$$\mathbb{E}(Y) = \frac{\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)}{n} = p$$
$$\text{var}(Y) = \frac{\text{var}(X_1) + \cdots + \text{var}(X_n)}{n^2} = \frac{p(1-p)}{n}$$
$$\text{stddev}(Y) = \sqrt{\frac{p(1-p)}{n}} \leq \frac{1}{2\sqrt{n}}$$

DSE 210: Probability and statistics **Winter 2018**

Worksheet 4 — Random variable, expectation, and variance

1. A die is thrown twice. Let X_1 and X_2 denote the outcomes, and define random variable X to be the minimum of X_1 and X_2. Determine the distribution of X.

2. A fair die is rolled repeatedly until a six is seen. What is the expected number of rolls?

3. On any given day, the probability it will be sunny is 0.8, the probability you will have a nice dinner is 0.25, and the probability that you will get to bed early is 0.5. Assume these three events are independent. What is the expected number of days before all three of them happen together?

4. An elevator operates in a building with 10 floors. One day, n people get into the elevator, and each of them chooses to go to a floor selected uniformly at random from 1 to 10.

   (a) What is the probability that exactly one person gets out at the ith floor? Give your answer in terms of n.

   (b) What is the expected number of floors in which exactly one person gets out? *Hint:* let X_i be 1 if exactly one person gets out on floor i, and 0 otherwise. Then use linearity of expectation.

5. You throw m balls into n bins, each independently at random. Let X be the number of balls that end up in bin 1.

   (a) Let X_i be the event that the ith ball falls in bin 1. Write X as a function of the X_i.

   (b) What is the expected value of X?

6. There is a dormitory with n beds for n students. One night the power goes out, and because it is dark, each student gets into a bed chosen uniformly at random. What is the expected number of students who end up in their own bed?

7. In each of the following cases, say whether X and Y are independent.

   (a) You randomly permute $(1, 2, \ldots, n)$. X is the number in the first position and Y is the number in the second position.

   (b) You randomly pick a sentence out of *Hamlet*. X is the first word in the sentence and Y is the second word.

   (c) You randomly pick a card from a pack of 52 cards. X is 1 if the card is a nine, and is 0 otherwise. Y is 1 if the card is a heart, and is 0 otherwise.

   (d) You randomly deal a ten-card hand from a pack of 52 cards. X is 1 if the hand contains a nine, and is 0 otherwise. Y is 1 if *all* cards in the hand are hearts, and is 0 otherwise.

8. A die has six sides that come up with different probabilities:

$$\Pr(1) = \Pr(2) = \Pr(3) = \Pr(4) = 1/8, \ \Pr(5) = \Pr(6) = 1/4.$$

   (a) You roll the die; let Z be the outcome. What is $\mathbb{E}(Z)$ and $\mathrm{var}(Z)$?

(b) You roll the die 10 times, independently; let X be the *sum* of all the rolls. What is $\mathbb{E}(X)$ and $\text{var}(X)$?

(c) You roll the die n times and take the average of all the rolls; call this A. What is $\mathbb{E}(A)$? What is $\text{var}(A)$?

9. Let $X_1, X_2, \ldots, X_{100}$ be the outcomes of 100 independent rolls of a fair die.

   (a) What are $\mathbb{E}(X_1)$ and $\text{var}(X_1)$?

   (b) Define the random variable X to be $X_1 - X_2$. What are $\mathbb{E}(X)$ and $\text{var}(X)$?

   (c) Define the random variable Y to be $X_1 - 2X_2 + X_3$. What is $\mathbb{E}(Y)$ and $\text{var}(Y)$?

   (d) Define the random variable $Z = X_1 - X_2 + X_3 - X_4 + \cdots + X_{99} - X_{100}$. What are $\mathbb{E}(Z)$ and $\text{var}(Z)$?

10. Suppose you throw m balls into n bins, where $m \geq n$. For the following questions, give answers in terms of m and n.

    (a) Let X_i be the number of balls that fall into bin i. What is $\Pr(X_i = 0)$?

    (b) What is $\Pr(X_i = 1)$?

    (c) What is $\mathbb{E}(X_i)$?

    (d) What is $\text{var}(X_i)$?

11. Give an example of random variables X and Y such that $\text{var}(X + Y) \neq \text{var}(X) + \text{var}(Y)$.

12. Suppose a fair coin is tossed repeatedly until the same outcome occurs twice in a row (that is, two heads in a row or two tails in a row). What is the expected number of tosses?

13. In a sequence of coin tosses, a *run* is a series of consecutive heads or consecutive tails. For instance, the longest run in $HTHHHTTHHTHH$ consists of three heads. We are interested in the following question: when a fair coin is tossed n times, how long a run is the resulting sequence likely to contain?

    To study this, pick any k between 1 and n, and let R_k denote the number of runs of length exactly k (for instance, a run of length $k+1$ doesn't count). In order to figure out $\mathbb{E}(R_k)$, we define the following random variables: $X_i = 1$ if a run of length exactly k begins at position i, where $i \leq n - k + 1$.

    (a) What are $\mathbb{E}(X_1)$ and $\mathbb{E}(X_{n-k+1})$?

    (b) What is $\mathbb{E}(X_i)$ for $1 < i < n - k + 1$?

    (c) What is $\mathbb{E}(R_k)$?

    (d) What is, roughly, the largest k for which $\mathbb{E}(R_k) \geq 1$?

Modeling data with probability distributions

DSE 210

Distributional modeling

A useful way to summarize a data set:

- Fit a probability distribution to it.
- Simple and compact, and captures the big picture while smoothing out the wrinkles in the data.
- In subsequent application, use distribution as a proxy for the data.

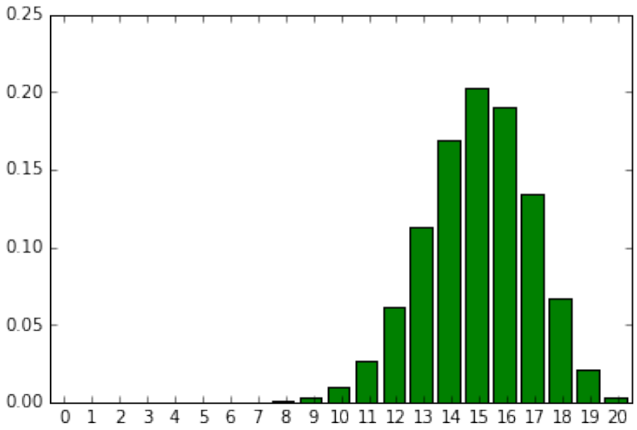
Which distributions to use?

> *There exist a few distributions of great universality which occur in a surprisingly large number of problems. The three principal distributions, with ramifications throughout probability theory, are the binomial distribution, the normal distribution, and the Poisson distribution.* – William Feller.

Well, this is true in one dimension. For higher-dimensional data, we'll use combinations of 1-d models: **products** and **mixtures**.

The binomial distribution

Binomial(n, p): the number of heads when n coins of bias (heads probability p) are tossed, independently.



Suppose X has a binomial(n, p) distribution.

$$\mathbb{E}X = np$$
$$\text{var}(X) = np(1-p)$$
$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Fitting a binomial distribution to data

Example: Upcoming election in a two-party country.

- You choose 1000 people at random and poll them.
- 600 say Democratic.

What is a good estimate for the fraction of votes the Democrats will get in the election? Clearly, 60%.

More generally, you observe n tosses of a coin of unknown bias. k of them are heads. How to estimate the bias?

$$p = \frac{k}{n}$$

Maximum likelihood estimation

Let $\mathcal{P}$ be a class of probability distributions (Gaussians, Poissons, etc).

Maximum likelihood principle: pick the distribution in $\mathcal{P}$ that makes the data maximally likely.
That is, pick the $p \in \mathcal{P}$ that maximizes $\Pr(\text{data}|p)$.

E.g. Suppose $\mathcal{P}$ is the class of binomials. We observe n coin tosses, and k of them are heads.

- Maximum likelihood : pick the bias p that maximizes

$$\Pr(\text{data}|p) = p^k(1-p)^{n-k}.$$

- Maximizing this is the same as maximizing its log,

$$\text{LL}(p) = k \ln p + (n-k)\ln(1-p).$$

- Set the derivative to zero.

$$\text{LL}'(p) = \frac{k}{p} - \frac{n-k}{1-p} = 0 \quad \Rightarrow \quad p = \frac{k}{n}.$$

Maximum likelihood: a small caveat

You have two coins of unknown bias.

- You toss the first coin 10 times, and it comes out heads every time. You estimate its bias as $p_1 = 1.0$.
- You toss the second coin 10 times, and it comes out heads once. You estimate its bias as $p_2 = 0.1$.

Now you are told that one of the coins was tossed 20 times and 19 of them came out heads. Which coin do you think it is?

- Likelihood under p_1:

$$\Pr(\text{19 heads out of 20 tosses}|\text{bias} = 1) = 0$$

- Likelihood under p_2:

$$\Pr(\text{19 heads out of 20 tosses}|\text{bias} = 0.1) = (0.1)^{19}(0.9)^1$$

The likelihood principle would choose the second coin. Is this right?

Laplace smoothing

A smoothed version of maximum-likelihood: when you toss a coin n times and observe k heads, estimate the bias as

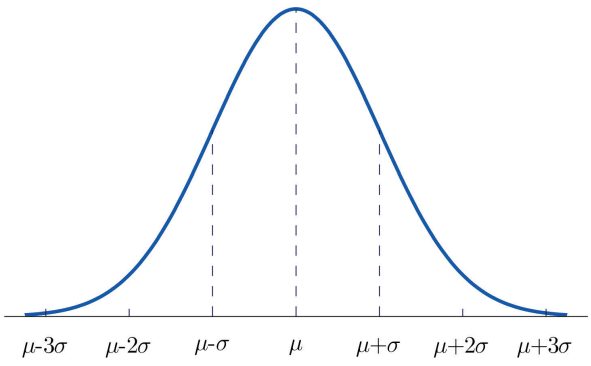
$$p = \frac{k+1}{n+2}.$$

Laplace's law of succession: What is the probability that the sun won't rise tomorrow?

- Let p be the probability that the sun won't rise on a randomly chosen day. We want to estimate p.
- For the past 5000 years ($= 1825000$ days), the sun has risen every day. Using Laplace smoothing, estimate

$$p = \frac{1}{1825002}.$$

The normal distribution



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

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

- 68.3% of the distribution lies within one standard deviation of the mean, i.e. in the range $\mu \pm \sigma$
- 95.5% lies within $\mu \pm 2\sigma$
- 99.7% lies within $\mu \pm 3\sigma$

Maximum likelihood estimation of the normal

Suppose you see n data points $x_1, \ldots, x_n \in \mathbb{R}$, and you want to fit a Gaussian $N(\mu, \sigma^2)$ to them. How to choose μ, σ?

- Maximum likelihood: pick μ, σ to maximize

$$\Pr(\text{data}|\mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2} \right) \right)$$

- Work with the log, since it makes things easier:

$$\text{LL}(\mu, \sigma^2) = \frac{n}{2} \ln \frac{1}{2\pi\sigma^2} - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.$$

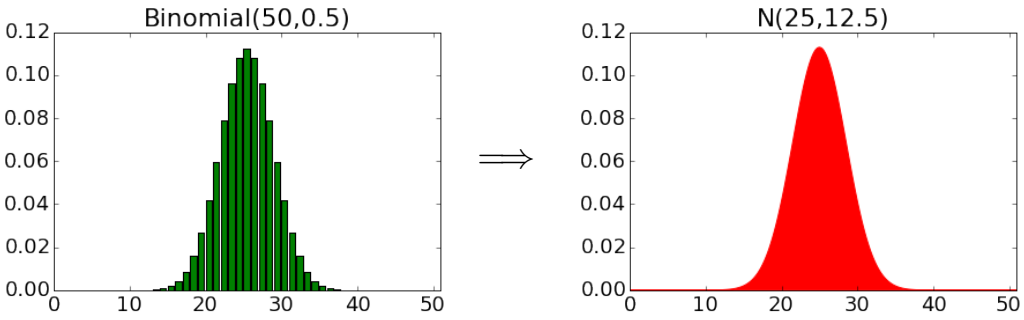
- Setting the derivatives to zero, we get

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

These are simply the empirical mean and variance.

Normal approximation to the binomial



When a coin of bias p is tossed n times, let X be the number of heads.

- We know X has mean np and variance $np(1-p)$.
- As n grows, the distribution of X looks increasingly like a Gaussian with this mean and variance.

Application to sampling

We want to find out what fraction p of San Diegans know how to surf. So we poll n random people, and find that k of them surf. Our estimate:

$$\widehat{p} = \frac{k}{n}.$$

Normal approximation:

- k has a binomial(n, p) distribution.
- This is close to a Gaussian with mean np and variance $np(1-p)$.
- Therefore the distribution of $\widehat{p} = k/n$ is close to a Gaussian with

$$\text{mean} = p$$
$$\text{variance} = \frac{p(1-p)}{n} \leq \frac{1}{4n}$$

Confidence intervals:

- With 95% confidence, our estimate is accurate within $\pm 1/\sqrt{n}$.
- With 99% confidence, our estimate is accurate within $\pm 3/2\sqrt{n}$.

The multinomial distribution

A k-sided die:

- A fair coin has two possible outcomes, each equally likely.
- A fair die has six possible outcomes, each equally likely.
- Imagine a k-faced die, with probabilities $p_1, \ldots, p_k$.

Toss such a die n times, and count the number of times each of the k faces occurs:

$$X_j = \#\text{ of times face } j \text{ occurs}$$

The distribution of $X = (X_1, \ldots, X_k)$ is called the **multinomial**.

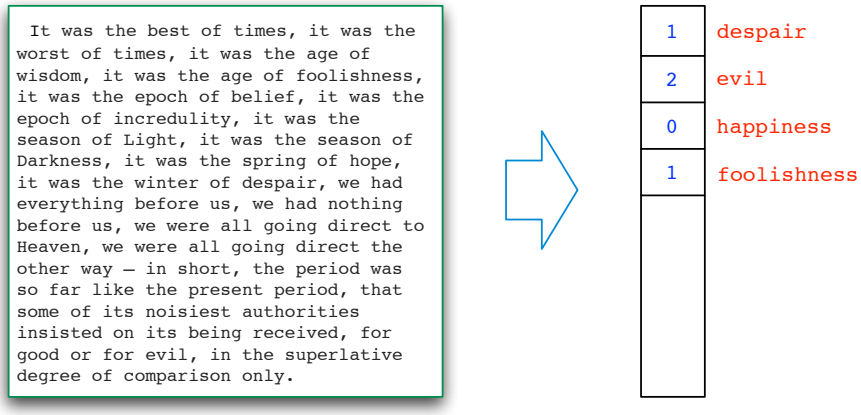
- Parameters: $p_1, \ldots, p_k \geq 0$, with $p_1 + \cdots + p_k = 1$.
- $\mathbb{E}X = (np_1, np_2, \ldots, np_k)$.
- $\Pr(n_1, \ldots, n_k) = \binom{n}{n_1, n_2, \ldots, n_k} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$, where

$$\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}$$

is the number of ways to place balls numbered $\{1, \ldots, n\}$ into bins numbered $\{1, \ldots, k\}$.

Example: text documents

Bag-of-words: vectorial representation of text documents.

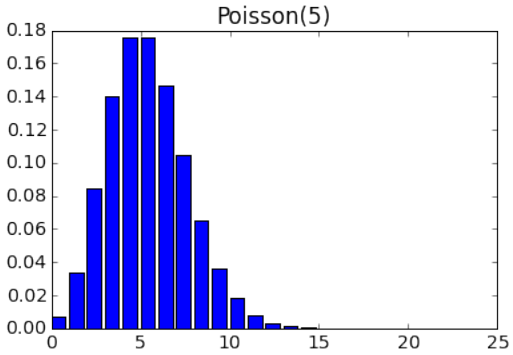


- Fix V = some vocabulary.
- Treat the words in a document as independent draws from a multinomial distribution over V:

$$p = (p_1, \ldots, p_{|V|}), \quad \text{such that} \quad p_i \geq 0 \text{ and } \sum_i p_i = 1$$

The Poisson distribution

A distribution over the non-negative integers $\{0, 1, 2, \ldots\}$



The Poisson has parameter $\lambda > 0$, with $\Pr(X = k) = e^{-\lambda} \dfrac{\lambda^k}{k!}$

- Mean: $\mathbb{E}X = \lambda$
- Variance: $\mathbb{E}(X - \lambda)^2 = \lambda$
- Maximum likelihood fit: set λ to the empirical mean

How the Poisson arises

Count the number of events (collisions, phone calls, etc) that occur in a certain interval of time. Call this number X, and say it has expected value λ.

Now suppose we divide the interval into small pieces of equal length.

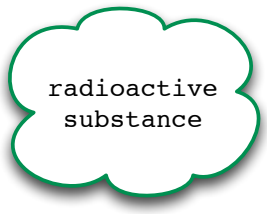
If the probability of an event occurring in a small interval is:

- independent of what happens in other small intervals, and
- the same across small intervals,

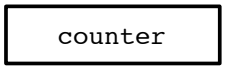
then $X \sim \text{Poisson}(\lambda)$.

Poisson: examples

Rutherford's experiments with radioactive disintegration (1920)

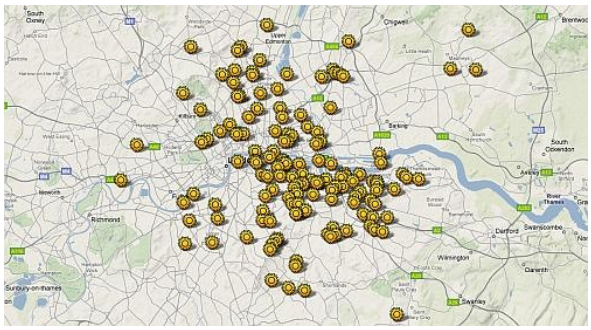


- $N = 2608$ intervals of 7.5 seconds
- $N_k = \#$ intervals with k particles
- Mean: 3.87 particles per interval



k	0	1	2	3	4	5	6	7	8	≥ 9
N_k	57	203	383	525	532	408	273	139	45	43
P(3.87)	54.4	211	407	526	508	394	254	140	67.9	46.3

Flying bomb hits on London in WWII



- Area divided into 576 regions, each 0.25 km^2
- $N_k = \#$ regions with k hits
- Mean: 0.93 hits per region

k	0	1	2	3	4	≥ 5
N_k	229	211	93	35	7	1
P(0.93)	226.8	211.4	98.54	30.62	7.14	1.57

Multivariate distributions

Almost all distributions we've considered are for one-dimensional data.

- Binomial, Poisson: integer
- Gaussian: real

What to do with the usual situation of data in higher dimensions?

1. Model each coordinate separately and treat them as independent.
   For $x = (x_1, \ldots, x_p)$, fit separate models $\Pr_i$ to each x_i, and assume

   $$\Pr(x_1, \ldots, x_p) = \Pr_1(x_1)\Pr_2(x_2)\cdots\Pr_p(x_p).$$

   This assumption is almost always completely inaccurate, and sometimes causes problems.
2. Multivariate Gaussian.
   Allows modeling of correlations between coordinates.
3. More general graphical models.
   Arbitrary dependencies between coordinates.

Classification with generative models 1

DSE 210

Machine learning versus Algorithms

In both fields, the goal is to develop

procedures that exhibit a desired input-output behavior.

- **Algorithms**: the input-output mapping can be precisely defined.
  Input: Graph G.
  Output: MST of G.
- **Machine learning**: the mapping cannot easily be made precise.
  Input: Picture of an animal.
  Output: Name of the animal.

Instead, we simply provide examples of (input,output) pairs and ask the machine to *learn* a suitable mapping itself.

Inputs and outputs

Basic terminology:

- The input space, $\mathcal{X}$.
  E.g. 32×32 RGB images of animals.
- The output space, $\mathcal{Y}$.
  E.g. Names of 100 animals.

x:

y: "bear"

After seeing a bunch of examples (x, y), pick a mapping

$$f : \mathcal{X} \to \mathcal{Y}$$

that accurately replicates the input-output pattern of the examples.

Learning problems are often categorized according to the type of *output space*: (1) discrete, (2) continuous, (3) probability values, or (4) more general structures.

Discrete output space: classification

Binary classification:

- Spam detection
  $\mathcal{X} = \{\text{email messages}\}$
  $\mathcal{Y} = \{\text{spam}, \text{not spam}\}$
- Credit card fraud detection
  $\mathcal{X} = \{\text{descriptions of credit card transactions}\}$
  $\mathcal{Y} = \{\text{fraudulent}, \text{legitimate}\}$

Multiclass classification:

- Animal recognition
  $\mathcal{X} = \{\text{animal pictures}\}$
  $\mathcal{Y} = \{\text{dog}, \text{cat}, \text{giraffe}, \ldots\}$
- News article classification
  $\mathcal{X} = \{\text{news articles}\}$
  $\mathcal{Y} = \{\text{politics}, \text{business}, \text{sports}, \ldots\}$

Continuous output space: regression

- Insurance company calculations
  What is the expected age until which this person will live?
  $\mathcal{Y} = [0, 120]$

- For the asthmatic
  Predict tomorrow's air quality (max over the whole day)
  $\mathcal{Y} = [0, \infty)$ (< 100: okay, > 200: dangerous)

What are suitable predictor variables ($\mathcal{X}$) in each case?

Conditional probability functions

Here $\mathcal{Y} = [0, 1]$ represents probabilities.

- Dating service
  What is the probability these two people will go on a date if introduced to each other?
  If we modeled this as a classification problem, the binary answer would basically always be "no". The goal is to find matches that are slightly less unlikely than others.
- Credit card transactions
  What is the probability that this transaction is fraudulent?
  The probability is important, because – in combination with the amount of the transaction – it determines the overall risk and thus the right course of action.

Structured output spaces

The output space consists of structured objects, like sequences or trees.

Dating service

Input: description of a person
Output: rank-ordered list of all possible matches

$\mathcal{Y}$ = space of all permutations

Example:
$x = Tom$
$y = (Nancy, Mary, Chloe, \ldots)$

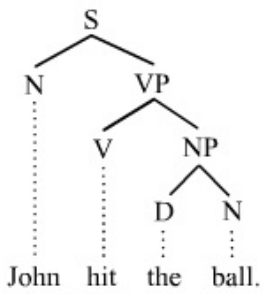
Language processing

Input: English sentence
Output: parse tree showing grammatical structure

$\mathcal{Y}$ = space of all trees

Example:
$x =$ "John hit the ball"

$y =$



A basic classifier: nearest neighbor

Given a labeled training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$.

Example: the MNIST data set of handwritten digits.

To classify a new instance x:

- Find its nearest neighbor amongst the $x^{(i)}$
- Return $y^{(i)}$

The data space

We need to choose a distance function.

Each image is 28×28 grayscale.
One option: Treat images as 784-dimensional vectors, and use Euclidean (ℓ_2) distance:

$$\|x - x'\| = \sqrt{\sum_{i=1}^{784} (x_i - x_i')^2}.$$

Summary:

- Data space $\mathcal{X} = \mathbb{R}^{784}$ with ℓ_2 distance
- Label space $\mathcal{Y} = \{0, 1, \ldots, 9\}$

Performance on MNIST

Training set of 60,000 points.

- What is the error rate on training points? **Zero.**
  In general, **training error** is an overly optimistic predictor of future performance.
- A better gauge: separate test set of 10,000 points.
  Test error = fraction of test points incorrectly classified.
- What test error would we expect for a random classifier? **90%.**
- Test error of nearest neighbor: **3.09%.**

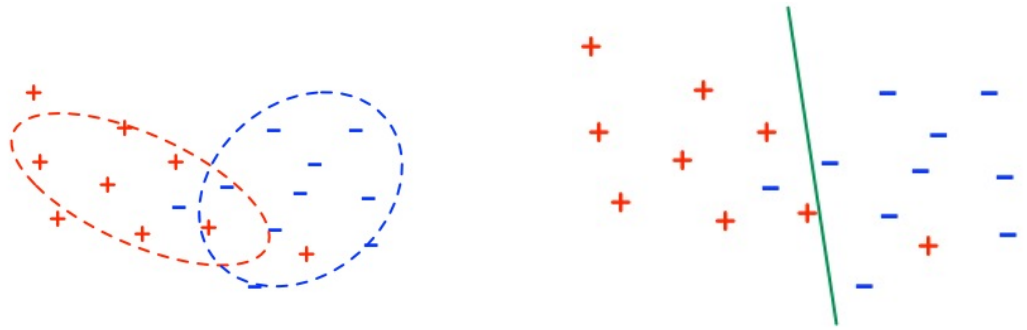
Examples of errors:

Properties of NN: (1) Can model arbitrarily complex functions
(2) Unbounded in size

Classification with parametrized models

Classifiers with a fixed number of parameters can represent a limited set of functions. Learning a model is about picking a good approximation.

Typically the x's are points in d-dimensional Euclidean space, $\mathbb{R}^d$.



Two ways to classify:

- **Generative**: model the individual classes.
- **Discriminative**: model the decision boundary between the classes.

Quick review of conditional probability

Formula for conditional probability: for any events A, B,

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

Applied twice, this yields Bayes' rule:

$$\Pr(H|E) = \frac{\Pr(E|H)}{\Pr(E)}\Pr(H).$$

Summation rule: Suppose events $A_1, \ldots, A_k$ are disjoint events, one of which must occur. Then for any other event E,

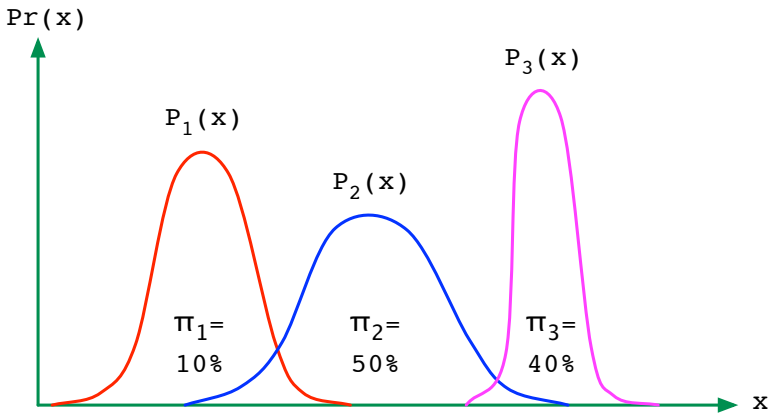
$$\begin{aligned}\Pr(E) &= \Pr(E, A_1) + \Pr(E, A_2) + \cdots + \Pr(E, A_k) \\ &= \Pr(E|A_1)\Pr(A_1) + \Pr(E|A_2)\Pr(A_2) + \cdots + \Pr(E|A_k)\Pr(A_k)\end{aligned}$$

Generative models

Generating a point (x, y) in two steps:

1. First choose y
2. Then choose x given y

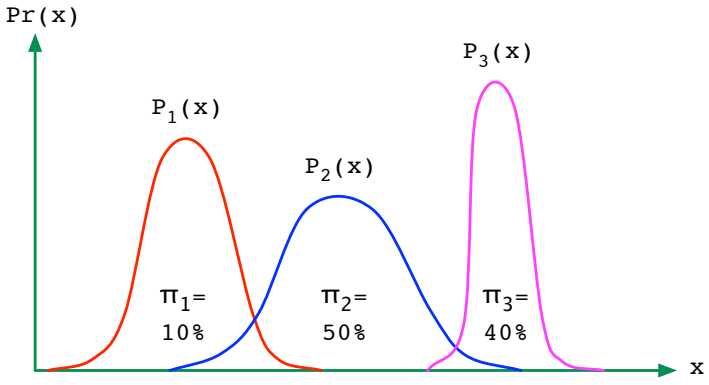
Example:
$\mathcal{X} = \mathbb{R}$
$\mathcal{Y} = \{1, 2, 3\}$



The overall density is a mixture of the individual densities,

$$\Pr(x) = \pi_1 P_1(x) + \cdots + \pi_k P_k(x).$$

The Bayes-optimal prediction



Labels $\mathcal{Y} = \{1, 2, \ldots, k\}$, density $\Pr(x) = \pi_1 P_1(x) + \cdots + \pi_k P_k(x)$.

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

$$\Pr(y = j|x) = \frac{\Pr(y = j)\Pr(x|y = j)}{\Pr(x)} = \frac{\pi_j P_j(x)}{\sum_{i=1}^{k} \pi_i P_i(x)}$$

Bayes-optimal (minimum-error) prediction: $h^*(x) = \arg\max_j \pi_j P_j(x)$.

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.

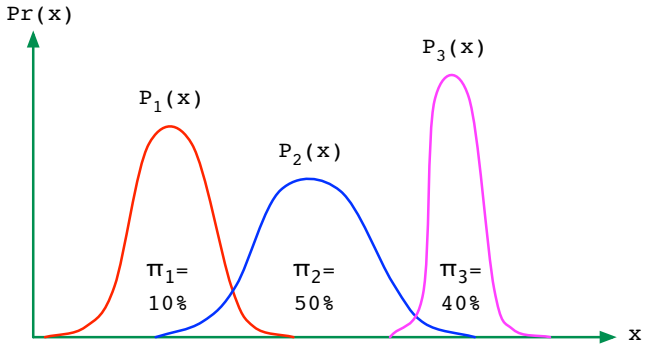
The data set

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

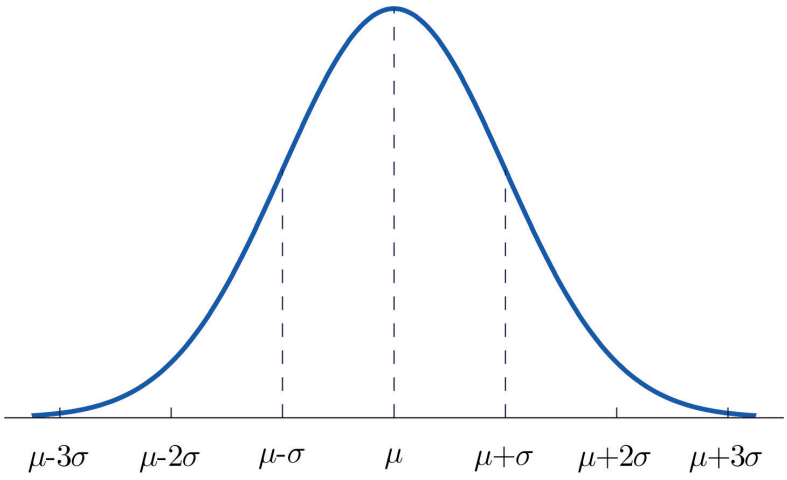
Training set of 130 bottles:

- 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

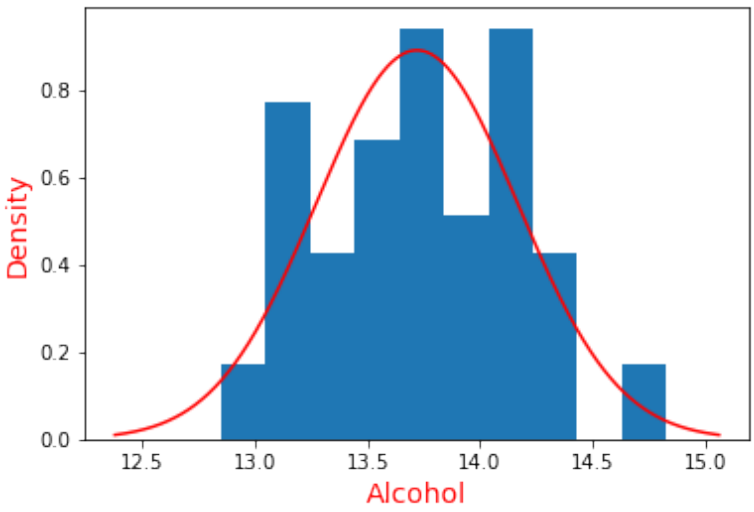


The Gaussian $N(\mu, \sigma^2)$ has mean μ, variance σ^2, and density function

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

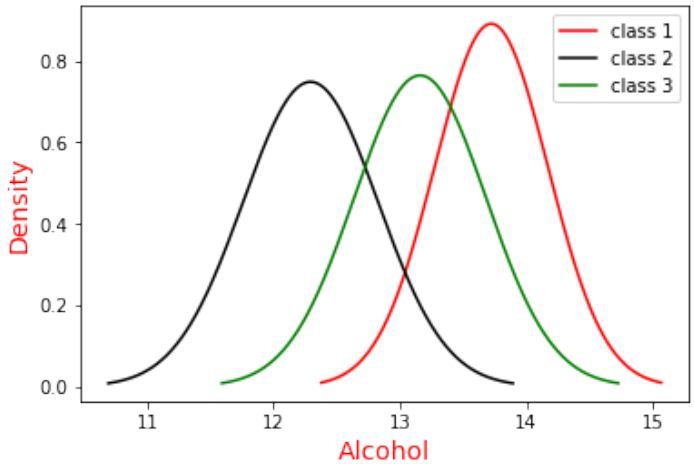
The distribution for winery 1

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%

DSE 210: Probability and statistics **Winter 2018**

Worksheet 5 — Classification with generative models 1

1. A man has two possible moods: `happy` and `sad`. The prior probabilities of these are:

$$\pi(\texttt{happy}) = \frac{3}{4}, \quad \pi(\texttt{sad}) = \frac{1}{4}.$$

His wife can usually judge his mood by how talkative he is. After much observation, she has noticed that:

- When he is happy,

$$\Pr(\text{talks a lot}) = \frac{2}{3}, \;\; \Pr(\text{talks a little}) = \frac{1}{6}, \;\; \Pr(\text{completely silent}) = \frac{1}{6}$$

- When he is sad,

$$\Pr(\text{talks a lot}) = \frac{1}{6}, \;\; \Pr(\text{talks a little}) = \frac{1}{6}, \;\; \Pr(\text{completely silent}) = \frac{2}{3}$$

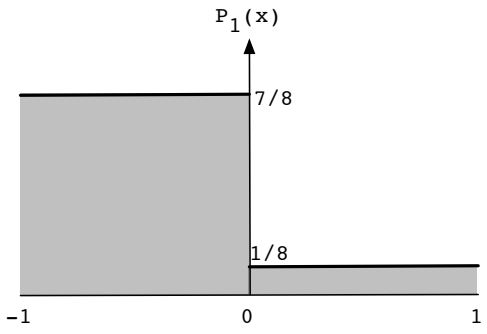
(a) Tonight, the man is just talking a little. What is his most likely mood?

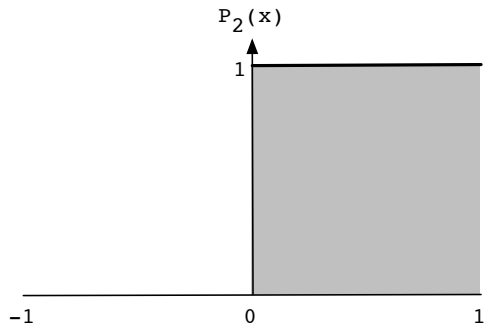
(b) What is the probability of the prediction in part (a) being incorrect?

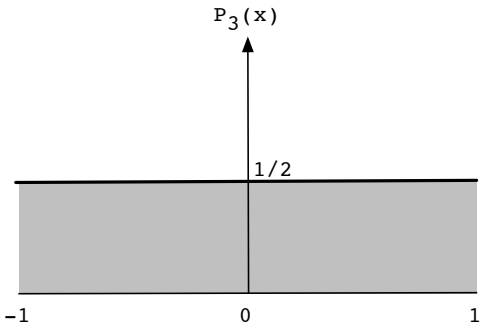
2. Suppose $\mathcal{X} = [-1, 1]$ and $\mathcal{Y} = \{1, 2, 3\}$, and that the individual classes have weights

$$\pi_1 = \frac{1}{3}, \;\; \pi_2 = \frac{1}{6}, \;\; \pi_3 = \frac{1}{2}$$

and densities P_1, P_2, P_3 as shown below.







What is the optimal classifier h^*? Specify it exactly, as a function from $\mathcal{X}$ to $\mathcal{Y}$.