Counting: permutations, combinations and the binomial theorem

Ten counting problems: the size of a hyperparameter grid and the number of ordered augmentation pipelines by the multiplication principle, choosing three features of ten against ranking them, splitting twelve examples into three minibatches with a multinomial coefficient, benchmark suites with at least one text task by the complement, the coefficient of x^2 in $(2x - 3)^5$ from the binomial theorem, counting feature subsets and even-sized subsets with $(1 + 1)^n$ and $(1 - 1)^n$, sharing ten GPU nodes among four jobs by stars and bars, the chance of a hash collision among thirty keys and the birthday bound, derangements of five shuffled labels by inclusion–exclusion, and drawing a test batch with exactly two positives by Vandermonde's identity, with worked solutions and the mistakes of counting arrangements when order does not matter, dropping the power on the coefficient inside a binomial term, choosing one item and then the rest, putting one bar too many in stars and bars, and adding overlapping cases instead of using inclusion–exclusion.

BEFORE YOU START

Most probability on a finite set of equally likely outcomes is counting: the probability of an event is the number of outcomes in it divided by the number of outcomes in all. The hard part is counting without listing, and almost every count comes from a handful of rules. Choices made one after another multiply. Arranging k of n distinct things in order gives a permutation; choosing them without caring about order divides that by the $k!$ orders and gives a combination $\binom{n}{k}$. The same $\binom{n}{k}$ are the coefficients of the binomial theorem, which is why the binomial distribution has them, and why sums of them can be found by substituting numbers into $(x + y)^n$. These ten problems count hyperparameter grids, feature subsets, minibatch splits, resource allocations and hash collisions, then use the counts as probabilities, including two cases where the obvious count is wrong and inclusion–exclusion fixes it.

- **Multiplication principle.** If a first choice can be made in a ways and, whatever it was, a second in b ways, the pair can be made in ab ways.
- **Permutations.** The number of ordered arrangements of k of n distinct items is $P(n, k) = n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$, and all n of them can be ordered in $n!$ ways, with $0! = 1$.
- **Combinations.** The number of k-element subsets of n items, order ignored, is $\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{P(n,k)}{k!}$. Symmetry: $\binom{n}{k} = \binom{n}{n-k}$. Pascal: $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.
- **Multinomial coefficients.** Splitting n distinct items into labeled groups of sizes $k_1, \ldots, k_m$ with $k_1 + \cdots + k_m = n$ can be done in $\frac{n!}{k_1!\cdots k_m!}$ ways. If groups of the same size are unlabeled, divide by the orders of those groups.

- **Binomial theorem.** $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$. Setting $x = y = 1$ gives $\sum_k \binom{n}{k} = 2^n$, and $x = -1$, $y = 1$ gives $\sum_k (-1)^k \binom{n}{k} = 0$ for $n \geq 1$.
- **Stars and bars.** The number of ways to write n as an ordered sum of k nonnegative integers, that is, to put n identical items into k distinct boxes, is $\binom{n+k-1}{k-1}$: n stars and $k-1$ bars. With every box nonempty it is $\binom{n-1}{k-1}$.
- **Complement and inclusion–exclusion.** "At least one" is often easiest as all minus none. For a union, $|A \cup B| = |A| + |B| - |A \cap B|$, and in general add the single sets, subtract the pairwise intersections, add the triples, and so on.

Problems

1. A grid search tries 4 learning rates, 3 batch sizes and 2 optimizers. How many configurations does it train? Separately, an augmentation pipeline applies 3 different transforms, chosen from 5, one after another, and the order matters. How many pipelines are there? difficulty ·

2. A model may keep 3 of 10 candidate features. How many different feature sets are possible? How many ways are there to rank the top 3 features, first, second and third? difficulty ·

3. Twelve distinct training examples are split into three minibatches of four. How many splits are there if the minibatches are labeled first, second and third? How many if only the grouping matters? difficulty ··

4. A benchmark suite of 4 tasks is drawn from 6 vision tasks and 5 text tasks. How many suites include at least one text task? If the suite is drawn at random, what is the probability that it does? difficulty ··

5. Find the coefficient of x^2 in $(2x-3)^5$, and the sum of all the coefficients of its expansion. difficulty ··

6. A model has 10 optional input features. How many subsets of the features are there, how many are nonempty, and how many have an even number of features? Use the binomial theorem. difficulty ··

7. A cluster has 10 identical GPU nodes to share among 4 jobs. How many allocations are there if a job may get no nodes? How many if every job gets at least one? difficulty ··

8. Thirty keys are hashed independently and uniformly into 1000 buckets. Find the probability that at least two keys share a bucket, compare it with the approximation $1 - e^{-n(n-1)/(2N)}$ and with the sum of the pair probabilities, and find the smallest number of keys for which a collision is more likely than not. difficulty ···

9. Five labeled test images are shuffled at random, so each permutation of the labels is equally likely. Find the number of permutations in which no image keeps its own label, and the probability of that. difficulty ···

10. A test set has 8 positive and 12 negative examples. A batch of 5 is drawn without replacement. Find the probability that exactly 2 positives are drawn, and use Vandermonde's identity to check that the probabilities for 0 to 5 positives add up to 1. difficulty ··

Answers

1. 24 configurations and 60 pipelines
2. 120 feature sets and 720 rankings
3. 34650 labeled splits and $34650/3! = 5775$ unlabeled ones
4. 315 suites, with probability $315/330 = 21/22 \approx 0.955$
5. The coefficient of x^2 is -1080; the coefficients sum to $(2-3)^5 = -1$
6. 1024 subsets, 1023 nonempty, and 512 of even size
7. 286 allocations, and 84 with every job busy
8. $\Pr(\text{collision}) \approx 0.356$; the approximation gives 0.353 and the pair sum 0.435. A collision becomes more likely than not at 38 keys
9. $120 - 76 = 44$ derangements, with probability $44/120 = 11/30 \approx 0.367$
10. $\Pr(2 \text{ positives}) = 6160/15504 \approx 0.397$, and the six probabilities add up to $15504/15504 = 1$

Worked solutions

Problem 1

> A grid search tries 4 learning rates, 3 batch sizes and 2 optimizers. How many configurations does it train? Separately, an augmentation pipeline applies 3 different transforms, chosen from 5, one after another, and the order matters. How many pipelines are there?

$4 \cdot 3 \cdot 2 = 24$	The multiplication principle: every learning rate pairs with every batch size and every optimizer.
5 choices for the first transform	Any of the five.
4 for the second, 3 for the third	A transform is not repeated, so each choice leaves one fewer.
$P(5,3) = 5 \cdot 4 \cdot 3$	Equivalently $\frac{5!}{2!}$.
24 configurations and 60 pipelines	If the order did not matter, the $3! = 6$ orders of each set of transforms would be one choice, and there would be $\binom{5}{3} = 10$.

Problem 2

> A model may keep 3 of 10 candidate features. How many different feature sets are possible? How many ways are there to rank the top 3 features, first, second and third?

$P(10,3) = 10 \cdot 9 \cdot 8 = 720$	Ordered choices: the ranking.
Each set of 3 features appears in $3! = 6$ rankings	The same three features ranked in every possible order.
$\binom{10}{3} = 720/6$	Divide out the orders.
$\binom{10}{3} = \frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1} = 120$	Check with symmetry: $\binom{10}{7}$ is also 120.

120 **feature sets and** 720 **rankings**	A feature set is a subset, so order does not matter; a ranking is an arrangement. Using 720 for the sets overcounts six times (Mistake 1).

Problem 3

> Twelve distinct training examples are split into three minibatches of four. How many splits are there if the minibatches are labeled first, second and third? How many if only the grouping matters?

$\binom{12}{4} = 495$ ways to fill the first batch	Choose its four examples.
$\binom{8}{4} = 70$ for the second, $\binom{4}{4} = 1$ for the third	Choose from what is left.
$495 \cdot 70 \cdot 1 = 34650$	The multiplication principle.
$\frac{12!}{4!\,4!\,4!} = 34650$	The same count as a multinomial coefficient.
34650 **labeled splits and** $34650/3! =$ 5775 **unlabeled ones**	Without labels, each grouping was counted once for each of the $3! = 6$ ways to number its three batches.

Problem 4

> A benchmark suite of 4 tasks is drawn from 6 vision tasks and 5 text tasks. How many suites include at least one text task? If the suite is drawn at random, what is the probability that it does?

$\binom{11}{4} = 330$ suites in all	Any 4 of the 11 tasks.
$\binom{6}{4} = 15$ suites with no text task	All four tasks from the six vision tasks.
$330 - 15 = 315$	The complement: all suites minus those with none.

Check: $\binom{5}{1}\binom{6}{3} + \binom{5}{2}\binom{6}{2} + \binom{5}{3}\binom{6}{1} + \binom{5}{4}$	Split by the number of text tasks, 1 to 4.
315 suites, with probability $315/330 = 21/22 \approx 0.955$	The check sums $100 + 150 + 60 + 5 = 315$. Picking one text task first and then any three of the rest gives 600 (Mistake 3).

Problem 5

> Find the coefficient of x^2 in $(2x - 3)^5$, and the sum of all the coefficients of its expansion.

$(2x - 3)^5 = \sum_k \binom{5}{k}(2x)^k(-3)^{5-k}$	The binomial theorem with the two terms $2x$ and -3, signs included.
$k = 2$: $\binom{5}{2}(2x)^2(-3)^3$	The term with x^2.
$\binom{5}{2} = 10$, $(2x)^2 = 4x^2$, $(-3)^3 = -27$	The 2 is squared along with the x.
$10 \cdot 4 \cdot (-27) = -1080$	Multiply the three factors.
The coefficient of x^2 is -1080; the coefficients sum to $(2 - 3)^5 = -1$	Setting $x = 1$ adds up every coefficient. Forgetting to square the 2 gives -540 (Mistake 2).

Problem 6

> A model has 10 optional input features. How many subsets of the features are there, how many are nonempty, and how many have an even number of features? Use the binomial theorem.

$\sum_{k=0}^{10} \binom{10}{k} = (1 + 1)^{10}$	Count the subsets of each size k and add; that is the binomial theorem at $x = y = 1$.
$2^{10} = 1024$	Equivalently, each feature is in or out: 2 choices, ten times.
$1024 - 1 = 1023$ nonempty	Remove the empty set.

$\sum_k (-1)^k \binom{10}{k} = (1-1)^{10} = 0$	So the even-sized and odd-sized subsets are equal in number.
1024 **subsets,** 1023 **nonempty, and** 512 **of even size**	Even plus odd is 1024 and even minus odd is 0, so each is 512. Directly: $1 + 45 + 210 + 210 + 45 + 1 = 512$.

Problem 7

> A cluster has 10 identical GPU nodes to share among 4 jobs. How many allocations are there if a job may get no nodes? How many if every job gets at least one?

An allocation is a row of 10 stars split by 3 bars	The stars before the first bar go to job 1, between the first and second bars to job 2, and so on.
13 positions, of which 3 are bars	Four jobs need only three dividers.
$\binom{13}{3} = 286$	Choose where the bars go.
At least one each: give every job a node first, then share the other 6	That leaves $\binom{6+3}{3} = \binom{9}{3}$.
286 **allocations, and** 84 **with every job busy**	$\binom{9}{3} = 84$, the same as $\binom{n-1}{k-1}$ with $n = 10$, $k = 4$. Using 4 bars for 4 jobs gives $\binom{14}{4} = 1001$ (Mistake 4).

Problem 8

> Thirty keys are hashed independently and uniformly into 1000 buckets. Find the probability that at least two keys share a bucket, compare it with the approximation $1 - e^{-n(n-1)/(2N)}$ and with the sum of the pair probabilities, and find the smallest number of keys for which a collision is more likely than not.

$\Pr(\text{no collision}) = \frac{1000 \cdot 999 \cdot \cdots 971}{1000^{30}}$	The i-th key must avoid the $i-1$ buckets already used: $P(1000, 30)$ ordered ways out of 1000^{30}.
$= \prod_{i=0}^{29} \left(1 - \frac{i}{1000}\right) \approx 0.6445$	The same product, one factor per key.

$1 - 0.6445 \approx 0.3555$	At least one collision, by the complement.
$\binom{30}{2} = 435$ pairs, each colliding with probability $\frac{1}{1000}$	So the approximation is $1 - e^{-0.435} \approx 0.3527$, and the sum of the pair probabilities is 0.435.
$\Pr(\text{collision}) \approx 0.356$; **the approximation gives** 0.353 **and the pair sum** 0.435. **A collision becomes more likely than not at** 38 **keys**	At 37 keys the probability is 0.4905, at 38 it is 0.5093; the rule of thumb $1.18\sqrt{N} \approx 37.3$ lands between. The pair sum overcounts the outcomes where several pairs collide, so it is only an upper bound.

Problem 9

> Five labeled test images are shuffled at random, so each permutation of the labels is equally likely. Find the number of permutations in which no image keeps its own label, and the probability of that.

A_i: image i keeps its label, $\lvert A_i \rvert = 4! = 24$	The other four labels can be in any order.
$\lvert A_i \cap A_j \rvert = 3! = 6$, triples $2! = 2$, quadruples $1! = 1$, all five $0! = 1$	Fixing more images leaves fewer labels to arrange.
$\lvert \bigcup A_i \rvert = 5 \cdot 24 - 10 \cdot 6 + 10 \cdot 2 - 5 \cdot 1 + 1$	Inclusion–exclusion, with $\binom{5}{j}$ ways to pick which j images are fixed.
$= 120 - 60 + 20 - 5 + 1 = 76$	Permutations with at least one image keeping its label.
$120 - 76 = 44$ **derangements, with probability** $44/120 = 11/30 \approx 0.367$	Close to $1/e \approx 0.368$, the limit as the number of images grows. Stopping after the first term gives 120 permutations with a fixed point and a probability of 0 (Mistake 5).

Problem 10

> A test set has 8 positive and 12 negative examples. A batch of 5 is drawn without replacement. Find the probability that exactly 2 positives are drawn, and use Vandermonde's identity to

> check that the probabilities for 0 to 5 positives add up to 1.

$\binom{20}{5} = 15504$ batches, all equally likely	Any 5 of the 20 examples.
$\binom{8}{2} = 28$ ways to pick the positives, $\binom{12}{3} = 220$ the negatives	Exactly two of one kind and three of the other.
$28 \cdot 220 = 6160$	The multiplication principle.
$\sum_{k=0}^{5} \binom{8}{k}\binom{12}{5-k} = \binom{20}{5}$	Vandermonde: a batch of 5 from 20 has some number k of positives, so counting by k counts every batch once.
$\Pr(2 \text{ positives}) = 6160/15504 \approx 0.397$, **and the six probabilities add up to** $15504/15504 = 1$	This is the hypergeometric distribution. Drawing with replacement would give the binomial $\binom{5}{2}(0.4)^2(0.6)^3 \approx 0.346$ instead.

Where this goes wrong

1. Counting arrangements when order does not matter

The feature sets of Problem 2 are counted by choosing the features one at a time.

10 choices for the first feature, 9 for the second, 8 for the third	Right so far: Problem 2.
"So there are $10 \cdot 9 \cdot 8$ feature sets."	That counts ordered triples. The set $\{a, b, c\}$ is counted as *abc*, *acb*, *bac*, *bca*, *cab* and *cba*.
720 feature sets ✗	There are $720/3! = 120$. Ask whether swapping two of the chosen items gives a different outcome: for a ranking it does, for a set it does not.

2. Dropping the power on the coefficient inside a binomial term

The term of $(2x - 3)^5$ with x^2 is written down from the pattern $\binom{5}{k}x^k$.

$\binom{5}{2}(2x)^2(-3)^3$	Right so far: Problem 5.
"$(2x)^2$ contributes $2x^2$."	The square applies to the whole term $2x$, so it is $4x^2$. The same goes for the sign: $(-3)^3$ is -27, not 27.
The coefficient is $10 \cdot 2 \cdot (-27) = -540$ ✗	It is $10 \cdot 4 \cdot (-27) = -1080$. A quick check is $x = 1$: the coefficients must sum to $(2 - 3)^5 = -1$, and that only works out with every power applied to the whole term.

3. Choosing one item first and then the rest

To make sure the benchmark suite of Problem 4 has a text task, one is picked first.

5 ways to pick a text task	Right so far.
"Then any 3 of the other 10 tasks: $\binom{10}{3} = 120$."	A suite with two text tasks is now counted twice, once for each of them as "the" first pick, and a suite with three of them three times.
$5 \cdot 120 = 600$ suites with at least one text task ✗	There are only 330 suites in all. The right count is $330 - \binom{6}{4} = 315$, by the complement.

4. One bar too many in stars and bars

Ten identical GPU nodes go to four jobs, and each job seems to need a bar.

10 stars, one for each node	Right so far: Problem 7.
"4 jobs, so 4 bars, and 14 symbols in a row."	Bars separate the jobs. Four jobs need three bars; a fourth bar makes a fifth box.

$\binom{14}{4} = 1001$ allocations ✗	That counts the ways to share ten nodes among five jobs. With 3 bars it is $\binom{13}{3} = 286$. Check on a small case: 2 nodes for 2 jobs gives $(2,0)$, $(1,1)$, $(0,2)$, which is $\binom{3}{1} = 3$, not $\binom{4}{2} = 6$.

5. Adding overlapping cases instead of using inclusion–exclusion

The shuffled labels of Problem 9 are counted by adding up the ways each image can keep its label.

$\lvert A_i \rvert = 4! = 24$ for each of the 5 images	Right so far: Problem 9.
"At least one image keeps its label in $5 \cdot 24 = 120$ permutations."	A permutation that fixes two images is in two of the sets and was counted twice. The identity permutation, which fixes all five, was counted five times.
No permutation leaves every image mislabeled, since $120 - 120 = 0$ ✗	The overlaps must be subtracted and added back: $120 - 60 + 20 - 5 + 1 = 76$, which leaves 44 derangements, and a probability of $11/30$, not 0.

More at eulerpad.app. This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

EulerPad, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 1 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/counting-and-the-binomial-theorem