



Review of Probability

Applied Machine Learning

Derek Hoiem

What is a probability

- A belief, a confidence, a likelihood
- “There’s a 60% chance it will rain tomorrow.”
 - Based on the information I have, if we were to simulate the future 100 times, I’d expect it to rain 60 of them.
 - I think it’s a little more likely to rain than not
- You have a $1/18$ chance of rolling a 3 with two dice.
 - If you roll an infinite number of pairs of dice, 1 out of 18 of them will sum to 3.
- Probabilities are expectations, according to some information and assumptions.
 - E.g., it will either rain tomorrow or not

Why do we care about probability in machine learning?

- ML problems are often formulated as maximizing a conditional probability $P(y|\mathbf{X})$, e.g. the probability of the true label given features, or maximizing the data likelihood $P(\mathbf{X})$
- Algorithms involving probabilistic objectives include logistic regression, naïve Bayes, decision trees, boosting, random forests, deep networks, EM algorithm, and more

Example

There are two movies showing:
“Bumblebee”, with 40 attendees, and
“Apocalypse” with 60 attendees.

What is the probability that a random
movie-goer is watching Apocalypse?

$P(M=A)$

- Out of all events, what fraction satisfy the criterion

Movie	Attendees
Bumblebee	40
Apocalypse	60

100 total people

Of those 60 are watching
Apocalypse

So $P(M = A) = 60 / 100 = 0.6$

Joint Probability

Suppose we also know whether each movie-goer is a child or adult

What is the probability that a movie-goer watches Bumblebee and is an adult?

- Out of all events, what fraction satisfy all criteria

Movie	Adult	Child
Bumblebee	20	20
Apocalypse	50	10

$$P(M = B, Age = Adult) = 20 / (20 + 20 + 50 + 10) = 0.2$$

Conditional Probability

Given that a movie-goer is a child, what is the probability that he or she is watching Bumblebee?

- Out of all events that satisfy the condition, what fraction satisfy the criterion

Movie	Adult	Child
Bumblebee	20	20
Apocalypse	50	10

$$P(M = B \mid Age = Child) = 20 / (20 + 10) = 0.667$$

Conditional Probability

If I know a movie-goer watching Bumblebee, what is the probability he or she is a child?

- Out of all events that satisfy the condition, what fraction satisfy the criterion

Movie	Adult	Child
Bumblebee	20	20
Apocalypse	50	10

$$P(M = B \mid \text{Age} = \text{Child}) = 20 / (20 + 20) = 0.5$$

Relationships between joint and conditional/marginal probabilities

Joint probability is the product of the conditional probability and the probability that the condition is true.

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

This extends to many variables with a chain rule.

$$P(X, Y, Z) = P(X|Y, Z)P(Y|Z)P(Z)$$

Marginalize out a variable by summing over its possible values

$$P(X = v) = \sum_i P(X = v, Y = y_i) = \sum_i P(X = v|Y = y_i)P(y = y_i)$$

$P(\text{Movie}, \text{Age})$

	Adult	Child
Bumblebee	0.2	0.2
Apocalypse	0.5	0.1

From the joint probability table, you can always compute probabilities of subsets of variables or conditional probabilities with those variables.

$$P(\text{Adult}) = 0.2 + 0.5$$

$$P(\text{Bumblebee}) = 0.2 + 0.2$$

$$P(\text{Adult} \mid \text{Bumblebee}) = 0.2 / (0.2 + 0.2)$$

A is independent of B if (and only if)

$$P(A, B) = P(A)P(B)$$

$$P(A|B) = P(A), \quad P(B|A) = P(B)$$

A and B are conditionally independent of C if (and only if)

$$P(A, B|C) = P(A|C)P(B|C)$$

E.g. whether you and I take an umbrella to school is independent given the weather.

Estimate a discrete probability function by counting

Movie	Age
B	C
B	A
A	A
A	A
B	A
A	A
...	

$$P(M = B) = \frac{1}{|N|} \sum_n \delta(M_n = B)$$

$$P(M = B, A = C) = \frac{1}{|N|} \sum_n \delta(M_n = B, A_n = C)$$

Probability density functions of continuous variables

What if the variable is continuous?

Precise values may never occur and be infinitesimally unlikely, e.g.

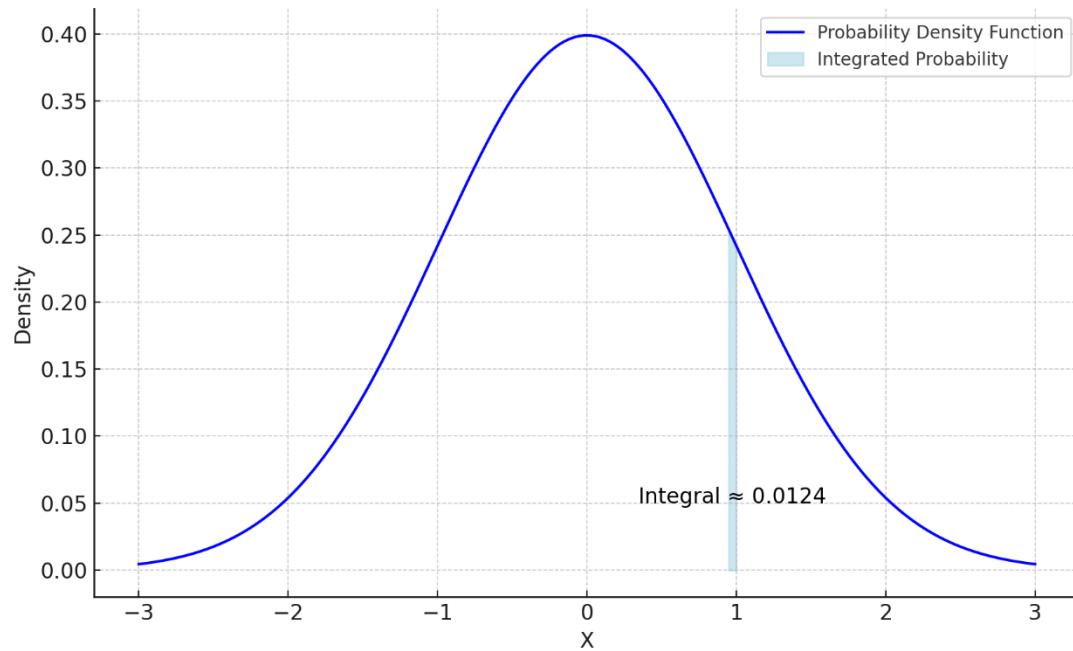
$$P(\text{Age} = 11.05) = 0$$

We replace the probability function with the probability *density* function, which can be integrated over a range to give a probability.

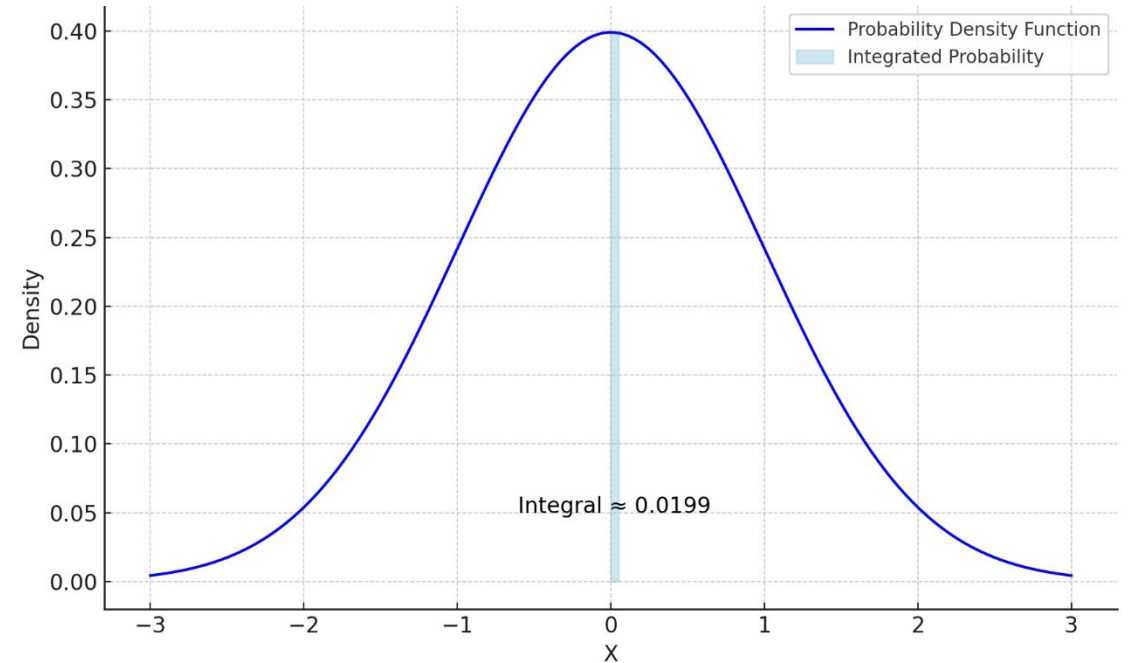
$$P(10 \leq \text{Age} < 11) = \int_{10}^{11} p(\text{Age} = a) da$$

Movie	Age
B	10
B	38
A	22
A	19
B	25
A	50
...	

Probability density function (PDF) → Probability



$$P(0.95 < X < 1) \approx 0.012$$



$$P(0 < X < 0.05) \approx 0.02$$

$$P(-0.05 < X < 0.05) \approx 0.04$$

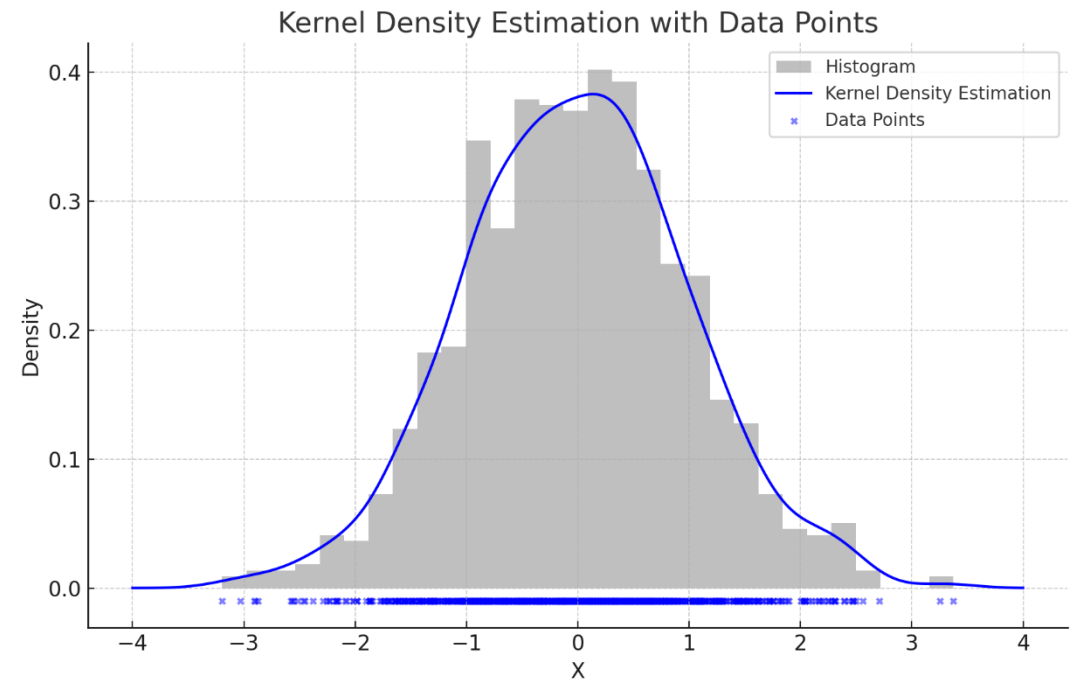
How to estimate probability density functions from samples

1. Discretize and count (histogram)

2. Kernel density estimation

3. Fit parameters of a model

Gaussian: $\mu = -0.015$ $\sigma = 1.004$



What must be true about probability functions?

1. Probabilities cannot be negative
2. The sum (for discrete) or integral (for continuous) must be 1
 - For discrete, this means that each value must be between 0 and 1
 - But values can be greater than 1 in a probability density function

Expectation and Variance

- The *expected value* or *mean* of a random variable $X \in \{v_1, \dots, v_N\}$ is the average value we'd get if we took an infinitely large sample:

$$E(X) = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{x_i \sim P(X)} x_i = \sum_{i=1..N} P(X = v_i) v_i$$

- We can take the expectation of a function $f(X)$:

$$E(f(X)) = \sum_{i=1..N} P(X = v_i) f(v_i)$$

- The *variance* of X measures the expected square difference of the values from the mean

$$\text{Var}(X) = E((E(X) - X)^2) = \sum_{i=1..N} P(X = v_i) (E(x) - v_i)^2$$

- We can also take the mean or variance of a sample $S = [s_0, \dots, s_K]$, called the empirical mean or variance

$$E(S) = \frac{1}{K} \sum_{i=1..K} s_i \quad \text{Var}(S) = \frac{1}{K} \sum_{i=1..K} (E(S) - s_i)^2$$

Entropy: a measure of expected surprise

- First, let's consider the likelihood of a set of values $\mathbf{x} = \{x_0, x_1, \dots\}$ s.t. $x_i \sim P(X)$: $P(\mathbf{x}) = \prod_i P(x_i)$, assuming the values are independent. However, this will become inconveniently small, so we can equivalently consider $\log P(\mathbf{x}) = \sum_i \log P(x_i)$. The expected value of this log likelihood gives us a measure of predictability.

- Entropy $H(X)$ is the negative expected log likelihood of variable $X \in \{v_1, \dots, v_N\}$

$$H(X) = - \sum_{i=1..N} P(X = v_i) \log_2 P(X = v_i)$$

- Greater entropy means that the value of X is less predictable
 - $H(X)$ ranges from 0 (all X have the same value) to $\log_2 N$ (all values of X have the same likelihood)
- Shannon's source coding theorem shows that the minimum expected number of bits required to encode a string of K i.i.d. values sampled from $P(X)$ is $KH(X)$
 - Complicated to prove, but think of each bit as dividing the possible values into two equally likely sets
 - E.g. let $P(X=1)=1/4$, $P(X=2)=1/4$, $P(X=3)=0$, $P(X=4)=1/2$. One bit can split ($[1,2]$, 4) and, if needed, a second bit can split between 1 and 2. To encode, $1 \rightarrow 00$, $2 \rightarrow 01$, $4 \rightarrow 1$. This requires 1.5 bits on average. $H(X) = 0.25 * 2 + 0.25 * 2 + 0.5 * 1 = 1.5$. 1-1-1-01-00-10-1 = 4,4,4,2,1,4,2

$x \sim P(X)$ means that x is drawn (randomly sampled) from the distribution $P(X)$

Typical machine learning problem: predict Y from X

- Given some features X , we want to predict target variable y , e.g.
 - X = email text and header; y = spam or not spam
 - X = meteorological data; y = next day's high temperature
 - X = image of a handwritten number; $y = 0, 1, \dots, \text{or } 9$
- We often frame this probabilistically
 - To predict, select $y^* = \operatorname{argmax}_y P(y|X)$, i.e. choose the y that is most likely given X
 - To train, optimize parameters that maximize the likelihood of the labels of the training data given the features of the training data

Basics of vector/matrix multiplication

- $\mathbf{w}^T \mathbf{x} = \mathbf{w} \cdot \mathbf{x} = \sum_i w_i x_i$
- Element (i, j) of $\mathbf{AB}$ is the dot product of the i th row of $\mathbf{A}$ with the j th column of $\mathbf{B}$
- $\mathbf{AB} \neq \mathbf{BA}$
- If $\mathbf{A}$ is $N \times M$ size matrix and $\mathbf{B}$ is $M \times K$, then $\mathbf{AB}$ is $N \times K$
- If $\mathbf{A}$ is $N \times M$ size matrix and $\mathbf{B}$ is $L \times K$ with $L \neq M$ then $\mathbf{A}$ cannot be multiplied by $\mathbf{B}$

Partial derivatives

$$\frac{\partial}{\partial w_i} \mathbf{w}^T \mathbf{x} = \frac{\partial}{\partial w_i} \sum_i w_i x_i = x_i$$

Classification by maximizing label likelihood

Suppose we want to predict a label $y_i \in \{-1, 1\}$ from an image X_i . Given a set of N training examples, solve for model parameters $\mathbf{w}$ to maximize $P(y_1 \dots y_n | X_1 \dots X_N)$. I.e., find the model parameters that make the training labels most likely, given the training features.

1. Assume each training sample is a sample from an identical and independent distribution (iid assumption): $P(y_1 \dots y_n | X_1 \dots X_N) = \prod_{i=1..N} P(y_i | X_i)$

2. Maximizing a product is hard because the derivative is complicated. Instead, we can maximize a sum of logs

$$\log \prod_{i=1..N} P(y_i | X_i) = \sum_{i=1..N} \log P(y_i | X_i)$$

3. We need a function (a.k.a. a model) to output the probability given the label. Let's use linear logistic regression

$$f(X_i, \mathbf{w}) = \mathbf{w}^T X_i = \log P(y_i = 1 | X_i) - \log P(y_i = -1 | X_i) = \log \frac{P(y_i = 1 | X_i)}{P(y_i = -1 | X_i)}$$

4. This is called a logistic score or logit. We can convert the logit to a probability:

$$\frac{1}{1 + \exp\left(-\log \frac{P(y_i = 1 | X_i)}{P(y_i = -1 | X_i)}\right)} = \frac{1}{1 + \frac{P(y_i = -1 | X_i)}{P(y_i = 1 | X_i)}} = \frac{P(y_i = 1 | X_i)}{P(y_i = 1 | X_i) + P(y_i = -1 | X_i)} = P(y_i = 1 | X_i)$$

5. This function $\sigma(x) = 1/(1 + \exp(-x))$ is called a sigmoid. So $\sigma(\mathbf{w}^T X_i) = P(y_i = 1 | X_i)$. Also, $\sigma(-\mathbf{w}^T X_i) = P(y_i = -1 | X_i)$.

6. Now we can write our objective in terms of parameters, image features, and labels:

$$\mathbf{w}_{opt} = \operatorname{argmax}_{\mathbf{w}} \sum_i \log(y_i \sigma(\mathbf{w}^T X_i)) = \operatorname{argmin}_{\mathbf{w}} \sum_i -\log(y_i \sigma(\mathbf{w}^T X_i))$$

7. The argmin expression is the "loss". We can optimize by taking the derivative of this expression wrt $\mathbf{w}$ and performing gradient descent.

Problems

https://us.prairielearn.com/pl/course_instance/190627/assessment/2574779