15-388/688 - Practical Data Science: Basic probability and statistics

J. Zico Kolter
Carnegie Mellon University
Fall 2016

Outline

Probability in data science

Basic rules of probability

Some common distributions

Maximum likelihood estimation

Naive Bayes

Machine learning via maximum likelihood estimation

Announcements

Additional information on tutorial posted to class web page

Tutorial check-in is due this Wednesday (no extensions except for special circumstances, but you can use late days, and see tutorial write-up for information on grading)

Final tutorial now due on 11/2, you can use max of 2 late days (so absolute deadline is 11/4)

Evaluation of other student tutorials due 11/9

You may still switch topics, as long as you can still submit check-in, but be warned that we may not be able to provide feedback

Announcements 10/19

Mid-way class survey was released on piazza: 50% response so far (going by HW3 submission counts)

We'll address this in more detail on Monday

But, we did want to address one very valid point of feedback: the HW uses library calls never discussed in class, and a lot of time is spent figuring out the APIs (this is the reality of data science, to some extent, but we can definitely do better)

To address this, we're going to have recitation sections for each of the HWs from now on, covering libraries used in the HW (more details soon)

Outline

Probability in data science

Basic rules of probability

Some common distributions

Maximum likelihood estimation

Naive Bayes

Machine learning via maximum likelihood estimation

Basic probability and statistics

Thus far, in our discussion of machine learning, we have largely avoided any talk of probability

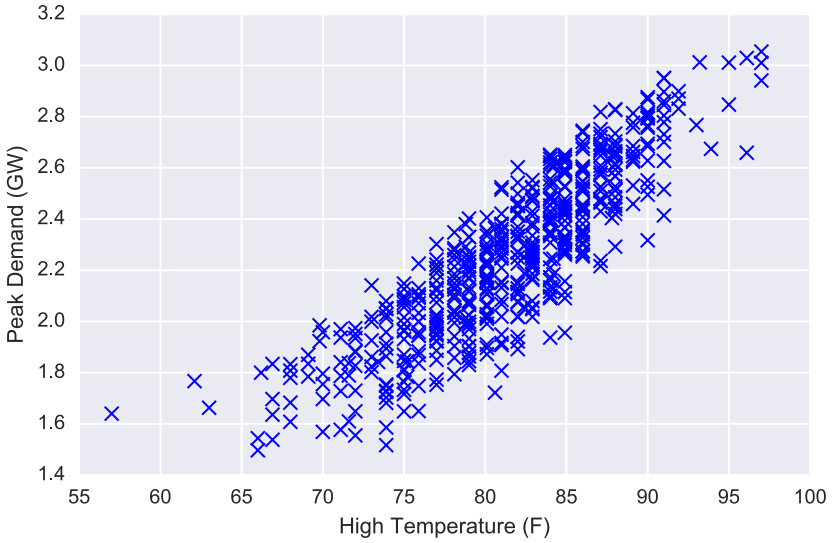
This won't be the case any longer, understanding and modeling probabilities is a crucial component of data science (and machine learning)

For the purposes of this course: statistics = probability + data

Probability and uncertainty in data science

In many prediction tasks, we never expect to be able to achieve perfect accuracy (there is some inherent randomness at the level we can observe the data)

In these situations, it is important to understand the uncertainty associated with our predictions



Outline

Probability in data science

Basic rules of probability

Some common distributions

Maximum likelihood estimation

Naive Bayes

Machine learning via maximum likelihood estimation

Random variables

A random variable (informally) is a variable whose value is not initial known

Instead, these variables can take on different values (including a possibly infinite number), and must take on exactly one of these values, each with an associated probability, which all together sum to one

$$\text{“Weather” takes values } \{\text{sunny}, \text{rainy}, \text{cloudy}, \text{snowy}\}$$
$$p(\text{Weather} = \text{ sunny}) = 0.3$$
$$p(\text{Weather} = \text{rainy}) = 0.2$$
$$\ldots$$

Slightly different notation for continuous random variables, which we will discuss shortly

Notation for random variables

In this lecture, we use upper case letters, X_i to denote random variables

For a random variable X_i taking values $\{1,2,3\}$

$$p(X_i) = \begin{pmatrix} 0.1 \\ 0.5 \\ 0.4 \end{pmatrix}$$

represents a set of probabilities for each value that X_i can take on (think of this like a dictionary mapping values of X_i) to numbers that sum to one

Conversely, we will use lower case x_i to denote a specific *value* of X_i (i.e., for above example $x_i \in \{1,2,3\}$), and $p(X_i = x_i)$ or just $p(x_i)$ refers to a *number* (the corresponding entry of $p(X_i)$)

Examples of probability notation

Given two random variables: X_1 with values in $\{1,2,3\}$ and X_2 with values in $\{1,2\}$:

$p(X_1, X_2)$ refers to the *joint distribution,* i.e., a set of 6 possible values for each setting of variables, i.e. a dictionary mapping $(1,1), (1,2), (2,1), \ldots$ to corresponding probabilities)

$p(x_1, x_2)$ is a *number:* probability that $X_1 = x_1$ and $X_2 = x_2$

$p(X_1, x_2)$ is a set of 3 values, the probabilities for all values of X_1 for the given value $X_2 = x_2$, i.e., it is a dictionary mapping $0,1,2$ to numbers (note: *not* probability distribution, it will not sum to one)

We generally call all of these terms ***factors*** (dictionaries mapping values to numbers, even if they do not sum to one)

Operations on probabilities/factors

We can perform operations on probabilities/factors by performing the operation on every corresponding value in the probabilities/factors

For example, given three random variables X_1, X_2, X_3:

$$p(X_1, X_2) \ \langle \text{op} \rangle \ p(X_2, X_3)$$

denotes a factor over X_1, X_2, X_3 (i.e., a dictionary over all possible combinations of values these three random variables can take), where the value for x_1, x_2, x_3 is given by

$$p(x_1, x_2) \ \langle \text{op} \rangle \ p(x_2, x_3)$$

Conditional probability

The ***conditional probability*** $p(X_1|X_2)$ (the conditional probability of X_1 given X_2) is defined as

$$p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

Can also be written $p(X_1, X_2) = p(X_1|X_2)p(X_2)$

More generally, leads to the ***chain rule:***

$$p(X_1, \dots, X_n) = \prod_{i=1}^{n} p(X_i|X_1, \dots X_{i-1})$$

Marginalization

For random variables X_1, X_2 with joint distribution $p(X_1, X_2)$

$$p(X_1) = \sum_{x_2} p(X_1, x_2) = \sum_{x_2} p(X_1|x_2)p(x_2)$$

Generalizes to joint distributions over multiple random variables

$$p(X_1, \dots, X_i) = \sum_{x_{i+1}, \dots, x_n} p(X_1, \dots, X_i, x_{i+1}, \dots, x_n)$$

For p to be a probability distribution, the marginalization over *all* variables must be one

$$\sum_{x_1, \dots, x_n} p(x_1, \dots, x_n) = 1$$

Bayes' rule

A straightforward manipulation of probabilities:

$$p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{\sum_{x_1} p(X_2|x_1)\, p(x_1)}$$

An example: I want to know if I have come with with a rate strain of value (occurring in only 1/10,000 people). There is an "accurate" test for the flu (if I have the flu, it will tell me I have 99% of the time, and if I do not have it, it will tell me I do not have it 99% of the time). I go to the doctor and test positive. What is the probability I have the this flu?

Independence

We say that random variables X_1 and X_2 are ***(marginally) independent*** if their joint distribution is the product of their marginals

$$p(X_1, X_2) = p(X_1)p(X_2)$$

Equivalently, can also be stated as the condition that

$$p(X_1|X_2) \quad \left(= \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X_1)p(X_2)}{p(X_2)} \right) = p(X_1)$$

$$(\text{and similarly}) \quad p(X_2|X_1) = p(X_2)$$

Conditional independence

We say that random variables X_1 and X_2 are ***conditionally independent given*** X_3, if

$$p(X_1, X_2|X_3) = p(X_1|X_3)p(X_2|X_3)$$

Again, can be equivalently written:

$$p(X_1|X_2, X_3) \left(= \frac{p(X_1, X_2|X_3)}{p(X_2|X_3)} = \frac{p(X_1|X_3)p(X_2|X_3)}{p(X_2|X_3)} \right)$$
$$= p(X_1|X_3)$$

And similarly $p(X_2|X_1, X_3) = p(X_2|X_3)$

Important: Marginal independence does not imply conditional independence or vice versa

Expectation

The expectation of a random variable is denoted:

$$\mathbf{E}[X] = \sum_x x \cdot p(x)$$

where we use upper case X to emphasize that this is a function of the entire random variable (but unlike $p(X)$ is a number)

Note that this only makes sense when the values that the random variable takes on are *numerical* (i.e., We can't ask for the expectation of the random variable "Weather")

Also generalizes to *conditional expectation:*

$$\mathbf{E}[X_1|x_2] = \sum_{x_1} x_1 \cdot p(x_1|x_2)$$

Rules of expectation

Expectation of sum is always equal to sum of expectations (even when variables are not independent):

$$\begin{aligned}\mathbf{E}[X_1 + X_2] &= \sum_{x_1,x_2} (x_1 + x_2) p(x_1, x_2) \\ &= \sum_{x_1} x_1 \sum_{x_2} p(x_1, x_2) + \sum_{x_2} x_2 \sum_{x_1} p(x_1, x_2) \\ &= \sum_{x_1} x_1 p(x_1) + \sum_{x_2} x_2 p(x_2) = \mathbf{E}[X_1] + \mathbf{E}[X_2]\end{aligned}$$

If x_1, x_2 independent, expectation of products is product of expectations

$$\begin{aligned}\mathbf{E}[X_1 X_2] &= \sum_{x_1,x_2} x_1 x_2 \, p(x_1, x_2) = \sum_{x_1,x_2} x_1 x_2 \, p(x_1) p(x_2) \\ &= \sum_{x_1} x_1 p(x_1) \sum_{x_2} x_2 p(x_2) = \mathbf{E}[X_1]\mathbf{E}[X_2]\end{aligned}$$

Variance

Variance of a random variable is the expectation of the variable minus its expectation, squared

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] \left(= \sum_x (x - \mathbf{E}[x])^2 p(x) \right) \\ &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] = \mathbf{E}[X^2] - \mathbf{E}[X]^2\end{aligned}$$

Generalizes to covariance between two random variables

$$\begin{aligned}\mathbf{Cov}[X_1, X_2] &= \mathbf{E}[(X_1 - \mathbf{E}[X_1])(X_2 - \mathbf{E}[X_2])] \\ &= \mathbf{E}[X_1 X_2] - \mathbf{E}[X_1]\mathbf{E}[X_2]\end{aligned}$$

Infinite random variables

All the math above works the same for discrete random variables that can take on an infinite number of values (for those with some math background, I'm talking about *countably infinite* values here*)*

The only difference is that $p(X)$ (obviously) cannot be specified by an explicit dictionary mapping variable values to probabilities, need to specify a *function* that produces probabilities

To be a probability, we still must have $\sum_x p(x) = 1$

Example:

$$P(X = k) = \left(\frac{1}{2}\right)^k, \qquad k = 1, \ldots, \infty$$

Continuous random variables

For random variables taking on *continuous* values (we'll only consider real-valued distributions), we need some slightly different mechanisms

As with infinite discrete variables, the distribution $p(X)$ needs to be specified as a function: here is referred to as a ***probability density function*** (PDF) and it must *integrate* to one $\int_{\mathbb{R}} p(x)dx = 1$

For any interval (a, b), we have that $p(a \leq x \leq b) = \int_a^b p(x)dx$ (with similar generalization to multi-dimensional random variables)

Can also be specified by their ***cumulative distribution function*** (CDF), $F(a) = p(x \leq a) = \int_{\infty}^{a} p(x)$

Outline

Probability in data science

Basic rules of probability

Some common distributions

Maximum likelihood estimation

Naive Bayes

Machine learning via maximum likelihood estimation

Bernoulli distribution

A simple distribution over binary $\{0,1\}$ random variables

$$p(X = 1; \phi) = \phi, \qquad P(X = 0; \phi) = 1 - \phi$$

where $\phi \in [0,1]$ is the parameter that governs the distribution

Expectation is just $\mathbf{E}[x] = \phi$ (but not very common to refer to it this way, since this would imply that the $\{0,1\}$ terms are actual real-valued numbers)

Categorical distribution

This is the discrete distribution we've mainly considered so far, a distribute over finite discrete elements with each probability specified

Written generically as:

$$p(X = i; \phi) = \phi_i$$

where $\phi_1, \ldots \phi_k \in [0,1]$ are the parameters of the distribution (the probability of each random variable, must sum to one)

Note: we could actually parameterize just using $\phi_1, \ldots \phi_{k-1}$, since this would determine the last elements

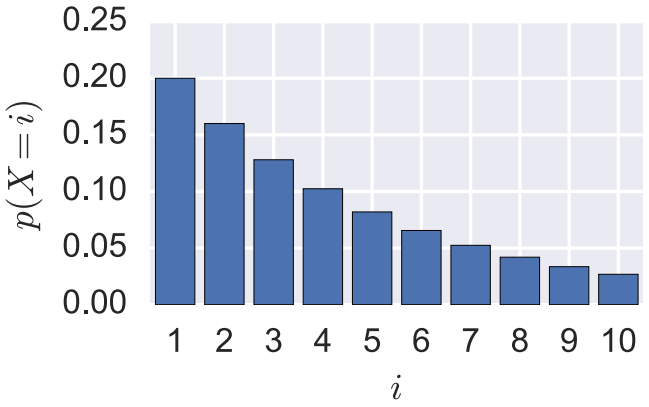
Unless the actual numerical value of the i's are relevant, it doesn't make sense to take expectations of a categorical random variable

Geometric distribution

The geometric distribution is an distribution over the positive integers, can be viewed as the number of Bernoulli trials needed before we get a "1"

$$p(X = i; \phi) = (1 - \phi)^{i-1}\phi, \qquad i = 1, \ldots, \infty$$

where $\phi \in [0,1]$ is parameter governing distribution (also $\mathbf{E}[X] = 1/\phi$)



$\phi = 0.2$

Note: easy to check that

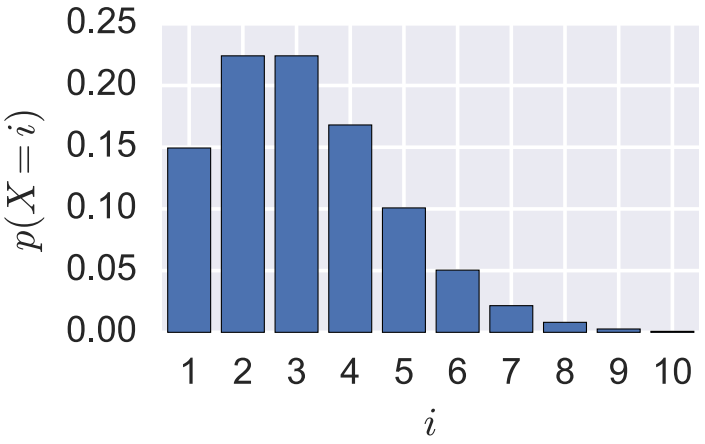
$$\sum_{i=1}^{\infty} p(X = i) = \phi \sum_{i=1}^{\infty} (1 - \phi)^{i-1} = \phi \cdot \frac{1}{1 - (1 - \phi)} = 1$$

Poisson distribution

Distribution over non-negative integers, popular for modeling number of times an event occurs within some interval

$$P(X = i; \lambda) = \frac{\lambda^i e^{-\lambda}}{i!}, \qquad i = 0, \ldots, \infty$$

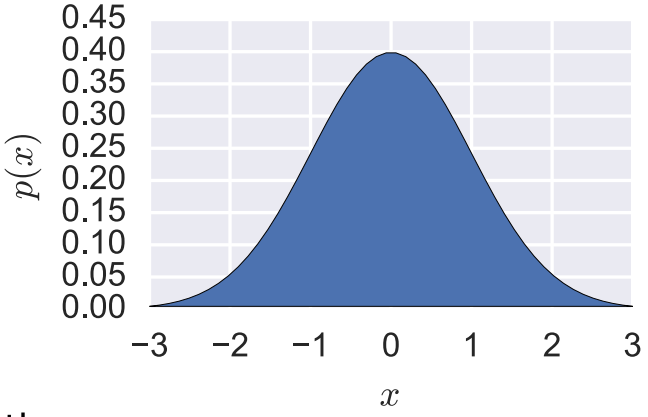
where $\lambda \in \mathbb{R}$ is parameter governing distribution (also $\mathbf{E}[X] = \lambda$)



$\lambda = 3$

Gaussian distribution

Distribution over real-valued numbers, empirically the most common distribution in all of data science (*not* in data itself, necessarily, but for people applying data science), the standard "bell curve":



$$\mu = 0$$
$$\sigma^2 = 1$$

Probability density function:

$$p(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \equiv \mathcal{N}(x; \mu, \sigma^2)$$

with parameters $\mu \in \mathbb{R}$ (mean) and $\sigma^2 \in \mathbb{R}_+$ (variance)

Multivariate Gaussians

The Gaussian distribution is one of the few distributions that generalizes nicely to higher dimensions

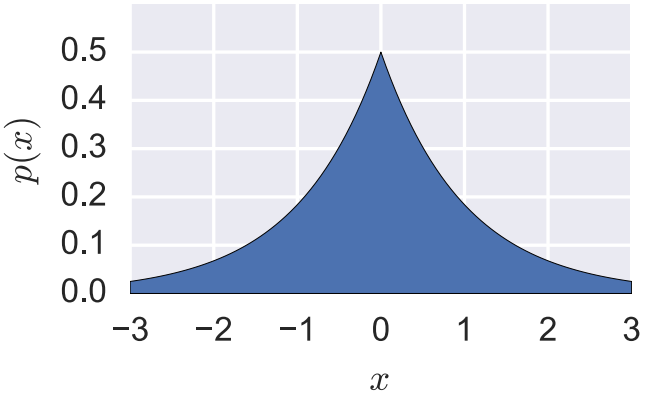
We'll discuss this in much more detail when we talk about anomaly detection and the mixture of Gaussians model, but for now, just know that we can also write a distribution over random *vectors* $x \in \mathbb{R}^n$

$$p(x; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\Big(-(x-\mu)^T \Sigma^{-1} (x-\mu)\Big)$$

where $\mu \in \mathbb{R}^n$ is mean and $\Sigma \in \mathbb{R}^{n \times n}$ is *covariance matrix,* and $|\cdot|$ denotes the determinant of a matrix

Laplace distribution

Like a Gaussian but with absolute instead of squared difference, gives the distribution (relatively) “heavy tails”



$\mu = 0$
$b = 1$

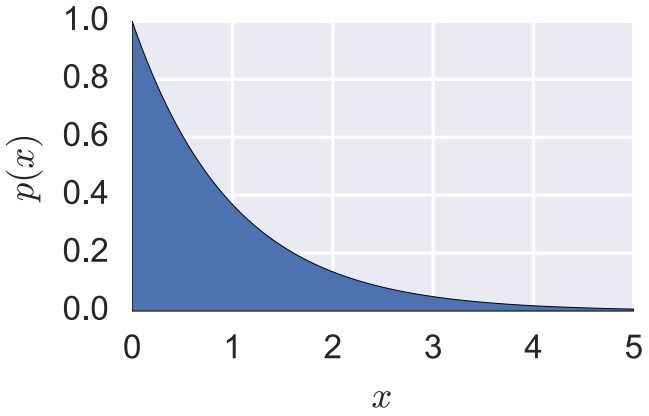
Probability density function:

$$p(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

with parameters μ (mean), b (variance is $2b^2$)

Exponential distribution

A one-sided Laplace distribution, often used to model arrival times



$\lambda = 1$

Probability density function:

$$p(x; \lambda) = \lambda \exp(-\lambda x)$$

with parameter $\lambda \in \mathbb{R}_+$ (mean/variance $\mathbf{E}[X] = 1/\lambda, \mathbf{Var}[x] = 1/\lambda^2$)

Some additional examples

Student's t distribution – distribution governing estimation of normal distribution from finite samples, commonly used in hypothesis testing

χ^2 (chi-squared) distribution – distribution of Gaussian variable squared, also used in hypothesis testing

Cauchy distribution – very heavy tailed distribution, to the point that variables have undefined expectation (the associated integral is undefined)

Outline

Probability in data science

Basic rules of probability

Some common distributions

Maximum likelihood estimation

Naive Bayes

Machine learning via maximum likelihood estimation

Estimating the parameters of distributions

We're moving now from probability to statistics

The basic question: given some data $x^{(1)}, \ldots, x^{(m)}$, how do I find a distribution that captures this data "well"?

In general (if we can pick from the space of all distributions), this is a hard question, but if we pick from a particular *parameterized family* of distributions $p(X; \theta)$, the question is (at least a little bit) easier

Question becomes: how do I find parameters θ of this distribution that fit the data?

Maximum likelihood estimation

Given a distribution $p(X;\theta)$, and a collection of observed (independent) data points $x^{(1)}, \ldots, x^{(m)}$, the probability of observing this data is simply

$$p(x^{(1)}, \ldots, x^{(m)}; \theta) = \prod_{i=1}^{m} p(x^{(i)}; \theta)$$

Basic idea of maximum likelihood estimation (MLE): find the parameters that maximize the probability of the observed data

$$\underset{\theta}{\text{maximize}} \prod_{i=1}^{m} p(x^{(i)}; \theta) \equiv \underset{\theta}{\text{maximize}} \ \ell(\theta) = \sum_{i=1}^{m} \log p(x^{(i)}; \theta)$$

where $\ell(\theta)$ is called the ***log likelihood*** of the data

Seems "obvious", but there are many other ways of fitting parameters

Parameter estimation for Bernoulli

Simple example: Bernoulli distribution

$$p(X = 1; \phi) = \phi, \qquad p(X = 0; \phi) = 1 - \phi$$

Given observed data $x^{(1)}, \ldots, x^{(m)}$, the "obvious" answer is:

$$\hat{\phi} = \frac{\#1\text{'s}}{\#\ \text{Total}} = \frac{\sum_{i=1}^{m} x^{(i)}}{m}$$

But why is this the case?

Maybe there are other estimates that are just as good, i.e.?

$$\phi = \frac{\sum_{i=1}^{m} x^{(i)} + 1}{m + 2}$$

MLE for Bernoulli

Maximum likelihood solution for Bernoulli given by

$$\underset{\phi}{\text{maximize}} \prod_{i=1}^{m} p(x^{(i)}; \phi) = \underset{\phi}{\text{maximize}} \prod_{i=1}^{m} \phi^{x^{(i)}} (1-\phi)^{1-x^{(i)}}$$

Taking the negative log of the optimization objective (just to be consistent with our usual notation of optimization as minimization)

$$\underset{\phi}{\text{maximize}}\ \ell(\phi) = \sum_{i=1}^{m} \left(x^{(i)} \log \phi + \left(1 - x^{(i)}\right) \log(1-\phi)\right)$$

Derivative with respect to ϕ is given by

$$\frac{d}{d\phi}\ell(\phi) = \sum_{i=1}^{m} \left(\frac{x^{(i)}}{\phi} - \frac{1-x^{(i)}}{1-\phi}\right) = \frac{\sum_{i=1}^{m} x^{(i)}}{\phi} - \frac{\sum_{i=1}^{m}(1-x^{(i)})}{1-\phi}$$

MLE for Bernoulli, continued

Setting derivative to zero gives:

$$\frac{\sum_{i=1}^{m} x^{(i)}}{\phi} - \frac{\sum_{i=1}^{m}(1 - x^{(i)})}{1-\phi} \equiv \frac{a}{\phi} - \frac{b}{1-\phi} = 0$$
$$\implies (1-\phi)a = \phi b$$
$$\implies \phi = \frac{a}{a+b} = \frac{\sum_{i=1}^{m} x^{(i)}}{m}$$

So, we have shown that the "natural" estimate of ϕ actually corresponds to the maximum likelihood estimate

MLE for Gaussian, briefly

For Gaussian distribution

$$p(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp(-(1/2)(x-\mu)^2/\sigma^2)$$

Log likelihood given by:

$$\ell(\mu, \sigma^2) = -m\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{m}\frac{(x^{(i)}-\mu)^2}{\sigma^2}$$

Derivatives (see if you can derive these fully):

$$\frac{d}{d\mu}\ell(\mu, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{m}\frac{x^{(i)}-\mu}{\sigma^2} = 0 \Longrightarrow \mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\frac{d}{d\sigma^2}\ell(\mu, \sigma^2) = -\frac{m}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{m}\frac{(x^{(i)}-\mu)^2}{(\sigma^2)^2} = 0 \Longrightarrow \sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu)^2$$

Outline

Probability in data science

Basic rules of probability

Some common distributions

Maximum likelihood estimation

Naive Bayes

Machine learning via maximum likelihood estimation

Naive Bayes modeling

Naive Bayes is a machine learning algorithm that rests relies heavily on probabilistic modeling

But, it is also interpretable according to the three ingredients of a machine learning algorithm (hypothesis function, loss, optimization), more on this later

Basic idea is that we model input and output as random variables $X = (X_1, X_2, ..., X_n)$ (several Bernoulli, categorical, or Gaussian random variables), and Y (one Bernoulli or categorical random variable), goal is to find $p(Y|X)$

Naive Bayes assumptions

We're going to find $p(Y|X)$ via Bayes' rule

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_y p(X|y)\,p(y)}$$

The denominator is just the sum over all values of Y of the distribution specified by the numeration, so we're just going to focus on the $p(X|Y)p(Y)$ term

Modeling full distribution $p(X|Y)$ for high-dimensional X is not practical, so we're going to make the ***naive Bayes assumption***, that the elements X_i are conditionally independent given Y

$$p(X|Y) = \prod_{i=1}^{n} p(X_i|Y)$$

Modeling individual distributions

We're going to explicitly model the distribution of each $p(X_i|Y)$ as well as $p(Y)$

We do this by specifying a distribution for $p(Y)$ and a *separate* distribution and for each $p(X_i|Y = y)$

So assuming, for instance, that Y_i and X_i are binary (Bernoulli random variables), then we would represent the distributions

$$p(Y; \phi_0), \qquad p(X_i|Y = 0; \phi_i^0), \qquad p(X_i|Y = 1; \phi_i^1)$$

We then estimate the parameters of these distributions using MLE, i.e.

$$\phi_0 = \frac{\sum_{j=1}^m y^{(j)}}{m}, \qquad \phi_i^y = \frac{\sum_{j=1}^m x_i^{(j)} \cdot 1\{y^{(j)} = y\}}{\sum_{j=1}^m 1\{y^{(j)} = y\}}$$

Making predictions

Given some new data point x, we can now compute the probability of each class

$$p(Y = y|x) \propto p(Y = y) \prod_{i=1}^{m} p(x_i|Y = y) = \phi_0 \prod_{i=1}^{m} (\phi_i^y)^{x_i} (1 - \phi_1^y)^{1-x_i}$$

After you have computed the right hand side, just normalize (divide by the sum over all y) to get the desired probability

Alternatively, if you just want to know the most likely Y, just compute each right hand side and take the maximum

Example

Y	X_1	X_2
0	0	0
1	1	0
0	0	1
1	1	1
1	1	0
0	1	0
1	0	1
?	1	0

$$p(Y = 1) = \ \phi_0 =$$

$$p(X_1 = 1|Y = 0) = \phi_1^0 =$$

$$p(X_1 = 1|Y = 1) = \phi_1^1 =$$

$$p(X_2 = 1|Y = 0) = \phi_2^0 =$$

$$p(X_2 = 1|Y = 0) = \phi_2^1 =$$

$$p(Y|X_1 = 1, X_2 = 0) =$$

Potential issues

Problem #1: when computing probability, the product $\mathrm{p}(y)\prod_{i=1}^{n} p(x_i|y)$ quickly goes to zero to numerical precision

Solution: compute log of the probabilities instead

$$\log p(y) + \sum_{i=1}^{n} \log p(x_i|y)$$

Problem #2: If we have never seen either $X_i = 1$ or $X_i = 0$ for a given y, then the corresponding probabilities computed by MLE will be zero

Solution: Laplace smoothing, "hallucinate" one $X_i = 0/1$ for each class

$$\phi_i^y = \frac{\sum_{j=1}^{m} x_i^{(j)} \cdot 1\{y^{(j)} = y\} + 1}{\sum_{j=1}^{m} 1\{y^{(j)} = y\} + 2}$$

Other distributions

Though naive Bayes is often presented as "just" counting, the value of the maximum likelihood interpretation is that it's clear how to model $p(X_i|Y)$ for non-categorical random variables

Example: if x_i is real-valued, we can model $p(X_i|Y=y)$ as a Gaussian

$$p(x_i|y;\mu^y,\sigma_y^2) = \mathcal{N}(x_i;\mu^y,\sigma_y^2)$$

with maximum likelihood estimates

$$\mu^y = \frac{\sum_{j=1}^m x_i^{(j)} \cdot 1\{y^{(j)}=y\}}{\sum_{j=1}^m 1\{y^{(j)}=y\}},\ \sigma_y^2 = \frac{\sum_{j=1}^m (x_i^{(j)} - \mu^y)\hat{}2 \cdot 1\{y^{(j)}=y\}}{\sum_{j=1}^m 1\{y^{(j)}=y\}}$$

All probability computations are exactly the same as before (it doesn't matter that some of the terms are probability densities)

Outline

Probability in data science

Basic rules of probability

Some common distributions

Maximum likelihood estimation

Naive Bayes

Machine learning via maximum likelihood estimation

Machine learning via maximum likelihood

Many machine learning algorithms (specifically the loss function component) can be interpreted probabilistically, as maximum likelihood estimation

Recall logistic regression:

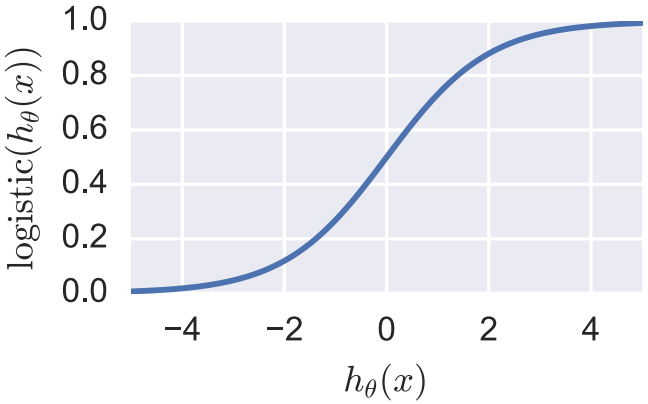
$$\underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \ell_{\text{logistic}}(h_\theta(x^{(i)}), y^{(i)})$$

$$\ell_{\text{logistic}}(h_\theta(x), y) = \log(1 + \exp(-y \cdot h_\theta(x)))$$

Logistic probability model

Consider the model (where Y is binary taking on $\{-1, +1\}$ values)

$$p(y|x;\theta) = \text{logistic}\big(y \cdot h_\theta(x)\big) = \frac{1}{1+\exp(-y \cdot h_\theta(x))}$$



Under this model, the maximum likelihood estimate is

$$\underset{\theta}{\text{maximize}} \sum_{i=1}^{m} \log p\big(y^{(i)}\big|x^{(i)};\theta\big) \equiv \underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \ell_{\text{logistic}}\big(h_\theta(x^{(i)}), y^{(i)}\big)$$

Least squares

In linear regression, assume

$$y = \theta^T x + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$
$$\iff p(y|x;\theta) = \mathcal{N}(\theta^T x, \sigma^2)$$

Then the maximum likelihood estimate is given by

$$\underset{\theta}{\text{maximize}} \sum_{i=1}^{m} \log p\big(y^{(i)}\big|x^{(i)};\theta\big) \equiv \underset{\theta}{\text{minimize}} \sum_{i=1}^{m} \big(y^{(i)} - \theta^T x^{(i)}\big)^2$$

i.e., the least-squares loss function can be viewed as MLE under Gaussian errors

Other approaches possible too: absolute loss function can be viewed as MLE under Laplace errors

Logistic regression vs. naive Bayes

Although we won't discuss it much more here, there is a very close connection between logistic regression and naive Bayes; for certain inputs we can show that both actually use the *same* hypothesis function

Logistic regression maximizes the conditional log likelihood (called a discriminative model)

$$\underset{\theta}{\text{maximize}} \sum_{i=1}^{m} \log p(y^{(i)} | x^{(i)}; \theta)$$

Naive Bayes maximizes the joint likelihood (called a *generative* model)

$$\underset{\theta}{\text{maximize}} \sum_{i=1}^{m} \log p(y^{(i)}, x^{(i)}; \theta)$$