Primer on Probabilities

Probability and Statistics Primer

- *Basic Concepts*
- *Maximum Likelihood Parameter Estimation*

Reading:

- *Many primers (check internet)*

e.g., Chapters 1,2 of

Pattern Recognition & Machine Learning by C. Bishop

A Probability Primer

- Assume an event where there is a degree of uncertainty in the outcome of the event
- **Random Variable:** A function which maps events or outcomes to a number set (i.e., integers, real etc)
- Refers to an event

$$R(\omega) = \begin{cases} 1 \text{ if } \omega = heads \\ 0 \text{ if } \omega = tails \end{cases}$$

Frequentistic Definition

- Frequency Probability:
  Probability $p(x)$ is the limit of its relative frequency in a large number of trials

$$p(x) = \lim_{n \to \infty} \frac{n_x}{n_t} \quad \text{or approx.} \quad p(x) \approx \frac{n_x}{n_t}$$

- It is the relative frequency with which an outcome would be obtained if the process were repeated a large number of times under exactly the same conditions.

Bayesian view

- Bayesian view: Probability is a measure of belief regarding the predicted outcome of an event.
- Uses the Bayes theorem to develop a calculus for performing probability reasoning.

Bayes theorem $$p(x,y) = p(x|y)p(y) \\ = p(y|x)p(x)$$

or: $$p(x|y) = \frac{p(x,y)}{p(y)} \qquad p(y|x) = \frac{p(x,y)}{p(x)}$$

or $$p(x|y) = p(y|x)\frac{p(x)}{p(y)}$$

Joint Probability Distribution

- Joint probabilities can be between any number of variables

  eg. $p(a=1, b=1, c=1)$
- For every combination of variables we need to define how probable that combination is
- The probabilities of all combinations need to sum up to 1.
- For 3 random variables taking two values the table contains 8 entries

a	b	c	$p(a,b,c)$
0	0	0	0.1
0	0	1	0.2
0	1	0	0.05
0	1	1	0.05
1	0	0	0.3
1	0	1	0.1
1	1	0	0.05
1	1	1	0.15

Sum up to 1

Joint Probability Distribution

a	b	c	$p(a,b,c)$
0	0	0	0.1
0	0	1	0.2
0	1	0	0.05
0	1	1	0.05
1	0	0	0.3
1	0	1	0.1
1	1	0	0.05
1	1	1	0.15

- Given the joint probability distribution, you can calculate any probability involving a, b, and c
- Note: May need to use marginalization and Bayes rule, (both of which are not discussed in these slides)

Examples of things you can compute:

- $p(a=1) = \sum_{b,c} p(a=1,b,c) sum\ rows\ a=1$
- $p(a=1,b=1|c=1) =$

  $p(a=1,b=1,c=1)|p(c=1)\ (Bayes\ Theorem)$

Bayes Theorem: An example

- School with 60% boys and 40% girls as its students.
- The female students wear trousers or skirts in equal numbers;
- All boys wear trousers.

An observer sees a (random) student from a distance, and what the observer can see is that this student is wearing trousers.

- ✓ What is the probability this student is a girl? The correct answer can be computed using Bayes' theorem.

Bayes Theorem: An example

$$s \qquad\qquad x$$

$$boy = 01, \quad girl = 10 \qquad trousers = 01, \quad skirt = 10$$

$$p(s = 01) = 0.6$$

$$p(x = 01|\, s = 01) = 1$$

$$p(x = 10|\, s = 01) = 0$$

$$p(s = 10) = 0.4$$

$$p(x = 01|\, s = 10) = 0.5$$

$$p(x = 10|\, s = 10) = 0.5$$

Bayes Theorem: An example

An observer sees a (random) student from a distance, and what the observer can see is that this student is wearing trousers.

- What is the probability this student is a girl?
  - ✓ The requested probability is translated as:

$$p(\, s = 10 | x = 01)$$

Bayes Theorem: An example

- Bayes theorem: $$p(s|x) = p(x|s)\frac{p(s)}{p(x)}$$

$$p(s=10|x=01) = p(x=01|s=10)\frac{p(s=10)}{p(x=01)}$$

- What do we know? $$p(x=01|\, s=10) = 0.5$$ $$p(s=10) = 0.4$$

- What are we missing? $p(x=01)$

Bayes Theorem: An example

- What can we find it? By marginalization.

$$p(x) = \sum_{s=01,10} p(x,s) = p(x,s=01) + p(x,s=10)$$

- And Bayes again to find $p(x=01)$

$$p(x=01,s=01) = p(x=01|s=01)p(s=01)$$
$$=1\text{x}0.6 = 0.6$$

$$p(x=01,s=10) = p(x=01|s=10)p(s=10)$$
$$= 0.5\text{x}0.4=0.2$$

- Hence $p(x=01) = 0.8$ and $\boxed{p(s=10|x=01) = 0.25}$

Independence

How is independence useful?

- Suppose you have n coin flips and you want to calculate the joint distribution $p(c_1, \ldots, c_n)$
- If the coin flips are not independent, you need 2^n values in the table
- If the coin flips are independent, then

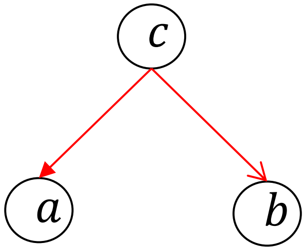
$$p(c_1, \ldots, c_n) = \prod_{i=1}^{n} p(c_i)$$

Each $p(c_i)$ table has 2 entries and there are n of them for a total of $2n$ values

Independence

Variables *a* and b are conditionally independent given *c* if any of the following hold:

- $p(a,b|c) = p(a|c)\, p(b|c)$
- $p(a|b,c) = p(a|c)$
- $p(b|a,c) = p(b|c)$

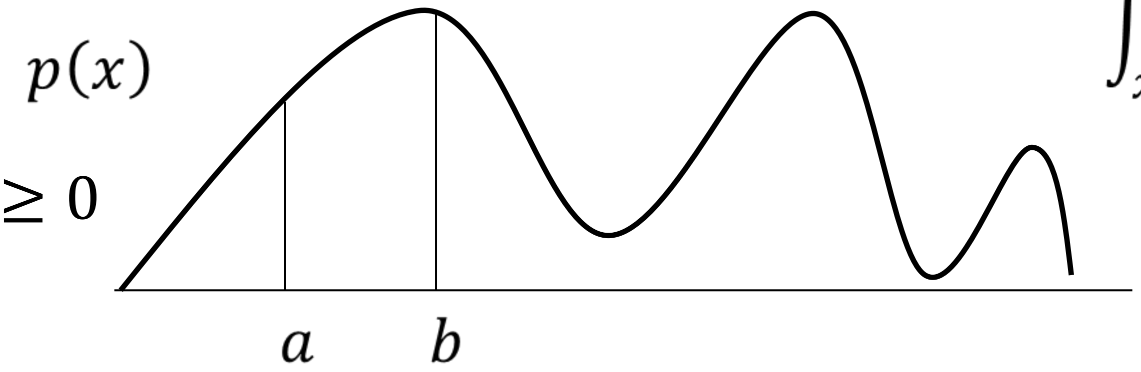


Knowing *c* tells me everything about *b* (ie. I don't gain anything by knowing *a)*. Either because *a* doesn't influence *b* or because knowing *c* provides all the information knowing *a* would give.

Continuous Variables

- Probability density function (PDF)

$$\int_x p(x)dx = 1$$



$p(x) \geq 0$

- Probability of $a < x < b$ $\quad P(a < x < b) = \int_{x=a}^{b} p(x)dx$

- Cumulative distribution function (CDF)

$$F(x < b) = \int_{x=-\infty}^{b} p(x)dx \qquad p(x) = \frac{dF(x)}{dx}$$

Continuous Variables

- Mean operator (first order moment)

$$E(x) = \int_x xp(x)dx$$

- Variance operator (second order moment)

$$E((x-\mu)^2) = \int_x (x-\mu)^2 p(x)dx$$

Popular PDFs

- Gaussian or Normal distribution

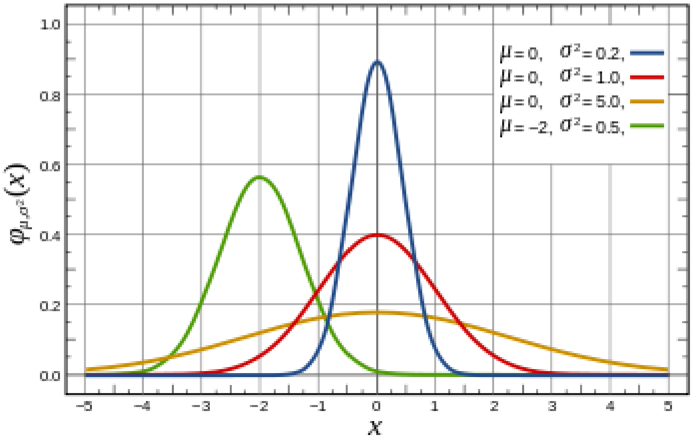
$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Parameters the mean and standard deviations μ, σ^2

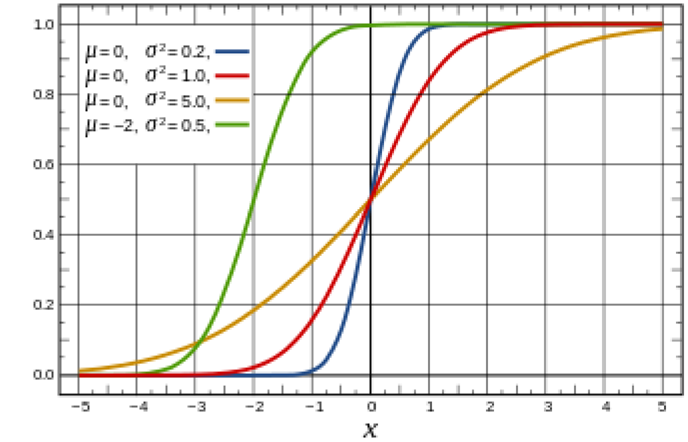
$$x \sim N(x|\mu, \sigma^2)$$

$$E(x) = \mu \qquad E\left((x-\mu)^2\right) = \sigma^2$$

- PDF



- CDF



Parameter estimation with Gaussians

- What is estimation? Given a set of observations and a model - estimate the model's parameter.

- First example: Given a population $\{x_1, x_2, x_3, \ldots x_N\}$ assuming that are independent samples from a normal distribution $x \sim N(x|\mu, \sigma^2)$ find an estimate for μ, σ^2

- How we approach the problem?

(1) We write the joint probability distribution (likelihood).

$$p(x_1, x_2, x_3, \ldots x_N | \mu, \sigma^2) = \prod_{i=1}^{N} p(x_i | \mu, \sigma^2)$$

Parameter estimation with Gaussians

(2) We substitute our distributional assumptions.

$$p(x_1, x_2,\ x_3, \ldots x_N|\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-\mu)^2/\sigma^2}$$

$$= \frac{(2\pi)^{-N/2}}{\sigma^N} e^{-\sum_{i=1}^{N}(x_i-\mu)^2/2\sigma^2}$$

(3) A common practice is to take the log of the joint function

$$\log p(\mu, \sigma^2) = -\frac{1}{2}\mathrm{N}\log 2\pi - \mathrm{N}\log\sigma - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}$$

Parameter estimation with Gaussians

(3) We take the maximum of the log likelihood (ML)

$$\mu_0, \sigma_0 = argmax_{\mu,\sigma} \log p(\mu, \sigma^2)$$

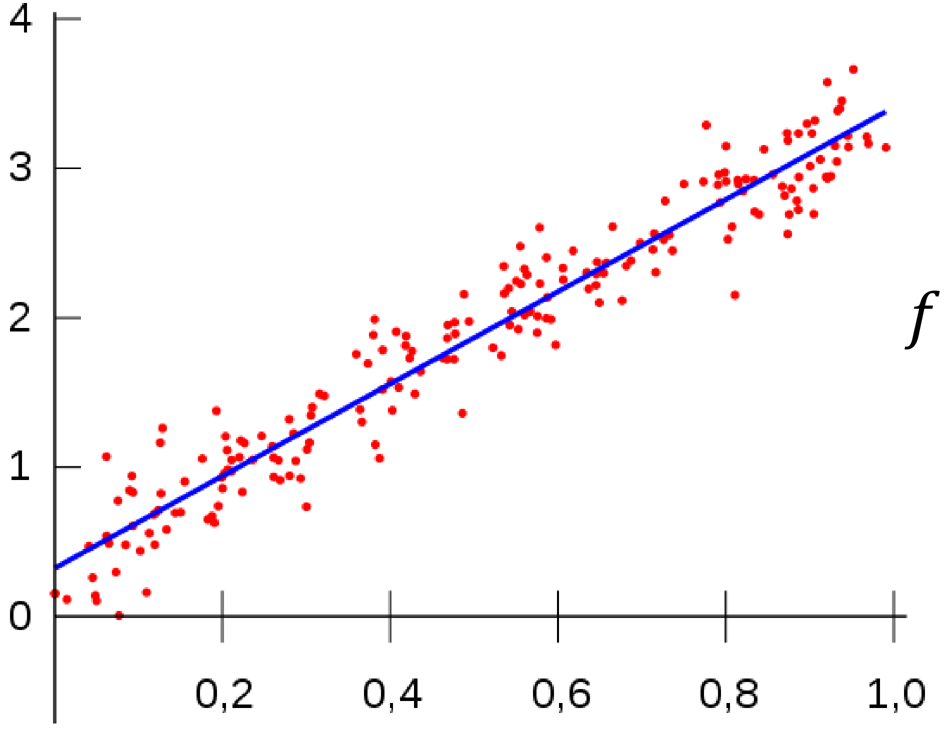
(4) We take the derivatives of p with regards to μ, σ^2

$$\frac{dlogp}{d\mu} = 0 \qquad \frac{dlogp}{d\sigma} = 0$$

$$\mu_0 = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \sigma_0 = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu_0)^2}{N}}$$

ML estimation: Linear Regression

The linear regression problem



y_i

x_i

$$f(x) = ax + b$$

ML estimation Linear Regression

Observations: $D = \{(x_1, y_1), \dots, (x_N, y_N)\}$

Model: $f(x) = ax + b$ and $y_i = f(x_i) + e_i$

$e_i \sim N(e_i|0, \sigma^2)$

Parameters: a, b, σ

Methodology: Maximum Likelihood

ML estimation Linear Regression

(1) We write the joint probability distribution (likelihood).

$$p((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) | f, \sigma^2)$$

$$= \prod_{i=1}^{N} p((x_i, y_i) | f, \sigma^2)$$

$$= \prod_{i=1}^{N} p(y_i | f, \sigma^2, x_i) \prod_{i=1}^{N} p(x_i)$$

we need to compute the following probability $p(y_i | f, \sigma^2, x_i)$

we know that

$$y_i = f(x_i) + e_i$$

$$e_i \sim N(e_i | 0, \sigma^2)$$

ML estimation Linear Regression

- Change of random variables

Assume a random variable a with pdf p_a

Assume a second random variable b and that a, b are related by $b = g(a)$

What is the pdf of b p_b ?

$$|p_b(b)db| = |p_a(a)da|$$

$$p_b(b) = |\frac{da}{db}|p_a(a)$$

$$p_b(b) = |\frac{dg^{-1}(b)}{db}|p_a(g^{-1}(b))$$

ML estimation Linear Regression

- Lets go back to our case

$$p(e_i) = N(e_i|0, \sigma^2) \quad y_i = g(e_i) = f(x_i) + e_i$$

$$e_i = y_i - f(x_i)$$

we compute $p(y_i)$ using the previous

$$p(y_i) = N(y_i - f(x_i)|0, \sigma^2)$$

or putting back what is constant we write more correctly

$$p(y_i|x_i, f, \sigma^2) = N(y_i|f(x_i), \sigma^2)$$

ML estimation: Linear Regression

$$p(D|f,\sigma^2) = \prod_{i=1}^{N} p(y_i|f,\sigma^2,x_i) \prod_{i=1}^{N} p(x_i)$$

$$= c \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - f(x_i))^2}$$

$$= c \frac{(2\pi)^{-N/2}}{\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - f(x_i))^2}$$

where c is constant $\mathrm{c} = \prod_{i=1}^{N} p(x_i)$

ML estimation: Linear Regression

- Choosing to maximize its logarithm we get

$$f^* = argmax_f - \mathrm{Nlog\ c2\pi} - \mathrm{Nlog}\ \sigma - \sum_{i=1}^{N} \frac{1}{2\pi\sigma^2}(y_i - f(x_i))^2$$

removing the constant terms we get

$$f^* = argmin_f \sum_{i=1}^{n} (f(x_i) - y_i))^2$$

ML estimation: Linear Regression

$$f^* = argmin_f \sum_{i=1}^{N}(ax_i + b - y_i)^2 \qquad g(a,b) = \sum_{i=1}^{N}(ax_i + b - y_i)^2$$

by taking the partial derivative with respect to a and b and setting them equal to zero we get

$$\frac{\partial g}{\partial b} = 0 \rightarrow 2\sum_{i=1}^{N}(ax_i + b - y_i) = 0 \rightarrow Nb = -\sum_{i=1}^{N} ax_i + \sum_{i=1}^{N} y_i \rightarrow$$

$$b = \bar{y} - a\bar{x} \ (1) \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad and \quad \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

ML estimation: Linear Regression

$$\frac{\partial g}{\partial a} = 0 \rightarrow x_i \sum_{i=1}^{N} (a x_i + b - y_i) = 0 \rightarrow \sum_{i=1}^{N} x_i y_i = a \sum_{i=1}^{N} {x_i}^2 + b \sum_{i=1}^{N} x_i$$

(2)

putting (1) into (2) we get

$$\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y} = a \sum_{i=1}^{N} {x_i}^2 - aN\,\bar{x}^2 \rightarrow a = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} {x_i}^2 - N\bar{x}^2}$$