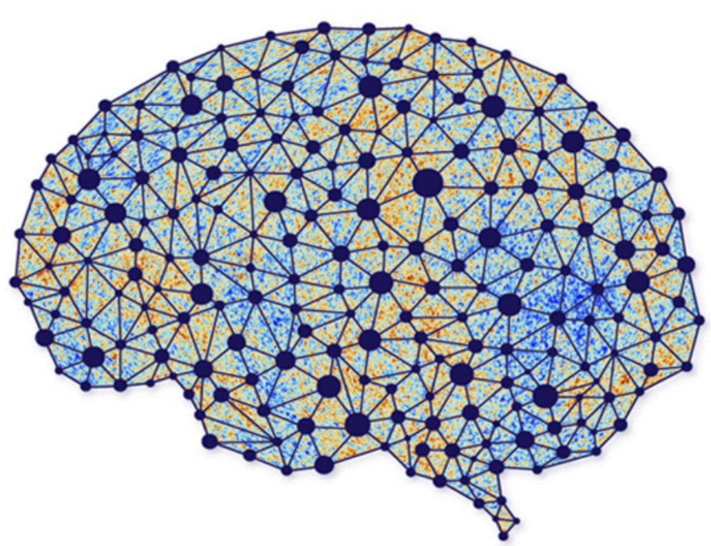
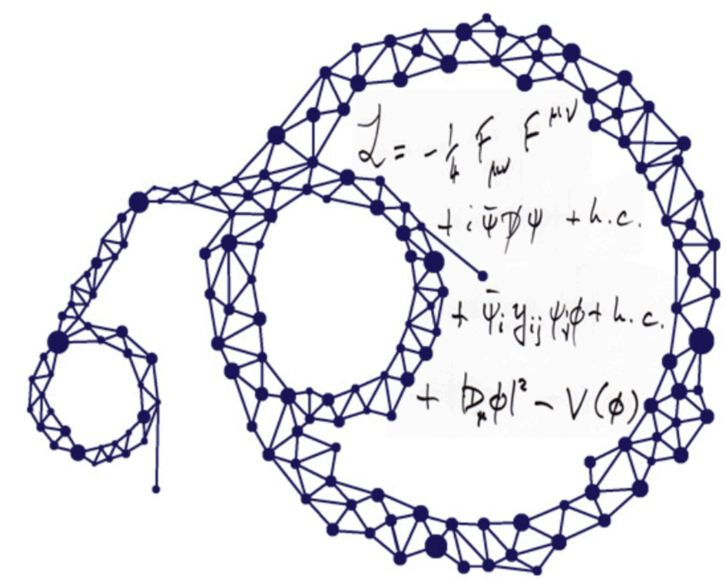
Physics 361 - Machine Learning in Physics

Lecture 2 – Background

Jan. 24th 2024



AI
∩
Universe



Moritz Münchmeyer

Unit 1: Background

Unit 1: Background

1.1 Probability Theory Background

Sources: e.g. deeplearningbook.org

Probability theory and Machine Learning

- **Data analysis in physics (and most domains) is always probabilistic**:
  - Inherent stochasticity in the system being modeled.
  - Incomplete observability.
  - Incomplete modeling.
- **Machine learning is inherently probabilistic**, and algorithms are written down using the notation of probability theory (e.g. expectation values).
- There are various forms of **probabilistic machine learning** that we will encounter, e.g.
  - Generative models represent PDFs
  - Machine learning of PDFs with normalizing flows, in Simulation-based Inference
  - Machine learning with probabilistic weights (Bayesian Neural Networks)
  - Machine learning can speed up more traditional statistical inference techniques such as MCMC.
- We will thus **frequently need concepts from probability theory** in this course.

Random variables

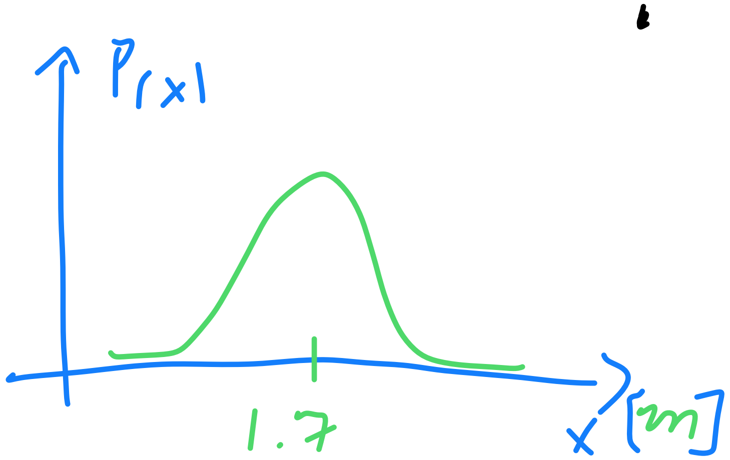
- A random variable x is sampled from
  - Probability density function $p(x)$ (continuous case)
  - Probability mass function $P(x)$ (discrete case)
- We often have vector valued random vars. $\vec{x}$
- For individual samples we write $x_1, x_2, x_3, \ldots$
- $x \sim P(x)$ means that x is sampled from $P(x)$
- [sometimes people write X : random variable
  x : sample of it]

For a continuous random var:

- $p(x) \geq 0$, but $p(x)$ can be >1

- $\int p(x)\, dx = 1$ normalization

$P(x = \text{height of a person})$



- To get a probability of an interval

$$P_{[a,b]} = \int_a^b p(x)\, dx$$

$$0 \leq P_{ab} \leq 1$$

For a discrete random variable

$P(x = \text{glasses})$

$$\begin{cases} x = \text{no gl.} & P(x) = 0.8 \\ x = \text{gl.} & P(x) = 0.2 \end{cases}$$

- $P(x) \leq 1$

- $\sum_x P(x) = 1$ normalization

Joint, conditional, marginal

- The Joint probability of two random vars. x and y is written as $P(x,y)$

↑ e.g. height and weight of a person

- The marginal probability is

$$p(x) = \int p(x,y)\, dy \qquad \text{(continous)}$$

$$P(x) = \sum_y P(x,y) \qquad \text{(discrete)}$$

- The conditional probability is

$$P(y|x) = \frac{P(x,y)}{P(x)}$$

„P of y given x"

Chain rule of conditional probability

given many random vars $x^{(1)}, x^{(2)} \dots$ we have

$$P(\mathrm{x}^{(1)}, \dots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \dots, \mathrm{x}^{(i-1)})$$

e.g.: $P(a, b, c) = P(a \mid b, c) \; P(b \mid c) \; P(c)$

Independence of random variables

$$P(x, y) = P(x) \, P(y) \quad \forall \, x, y$$

e.g. not the case for a person's weight and height

Expectation, Variance, Covariance

The expectation value of a funct. $f(x)$ of a random variable x is given by

$|\psi(x)|^2$ in QM

$$\mathbb{E}_{x\sim p}[f(x)] = \int p(x)\, f(x)\, dx \qquad \text{cont.}$$

$$\mathbb{E}_{x\sim p}[f(x)] = \sum_x P(x)\, f(x) \qquad \text{discrete}$$

In physics we like to write

$$\mathbb{E}[f(x)] \quad \text{as} \quad \langle f(x) \rangle$$

For example the mean: $\langle x \rangle = \int p(x)\, x\, dx$

Expectation values are Linear

$$\langle \alpha f(x) + \beta g(x) \rangle = \alpha \langle f(x) \rangle + \beta \langle g(x) \rangle$$

The Variance is defined as

$$\begin{aligned} \mathrm{Var}(f(x)) &= \langle (f(x) - \langle f(x) \rangle)^2 \rangle \\ &= \langle f^2 - 2f\langle f \rangle - \langle f \rangle^2 \rangle \\ &= \langle f^2 \rangle - 2\langle f \rangle^2 - \langle f \rangle^2 \\ &= \langle f^2 \rangle - \langle f \rangle^2 \end{aligned}$$

The standard deviation is

$$\sigma = \sqrt{\mathrm{Var}}$$

- The Covariance is

$$\mathrm{Cov}[f(x), g(y)] = \langle (f(x) - \langle f(x) \rangle)(g(y) - \langle g(y) \rangle) \rangle$$

function of random var x, e.g. height

function of random var y, e.g weight

- For a random vector $\vec{x}$ we have the covariance matrix

$$\mathrm{Cov}(\vec{x})_{i,j} = \mathrm{Cov}(x_i, x_j)$$

- and the correlation matrix is

$$\mathrm{Corr}(\vec{x})_{ij} = \frac{\mathrm{Cov}(\vec{x})_{ij}}{\sqrt{\mathrm{Cov}(\vec{x})_{ii}\,\mathrm{Cov}(\vec{x})_{jj}}} \in (-1, 1)$$

e.g $\mathrm{Corr}(\text{height}, \text{weight}) \sim 0.7$

Gaussian Distribution

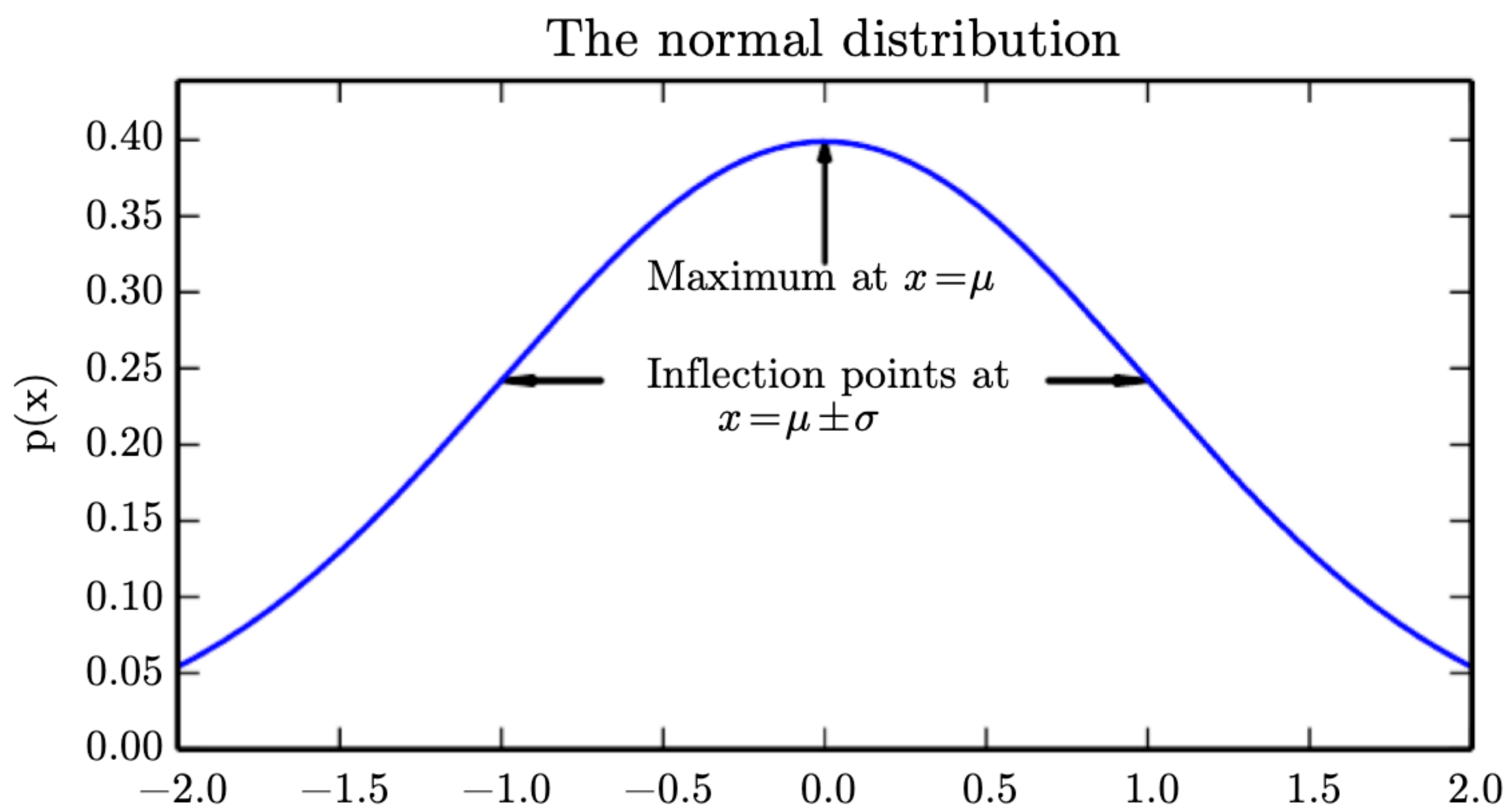
= „normal distribution"

mean

in 1 dim.: $\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

σ^2: variance

σ: standard dev.



deeplearningbook.org

covariance matrix

in N dim.: $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$

Bayes theorem

$$P(x,y) = P(y,x)$$

$$P(x|y)\,P(y) = P(y|x)\,P(x)$$

$$\Rightarrow \quad \boxed{P(x|y) = \frac{P(y|x)\,P(x)}{P(y)} \quad \text{Bayes theorem}}$$

In particular given $P(x|y)$ we can calculate $P(y|x)$ and vice versa.

Change of variables

If we have a random vector $\vec{x}$ and define a new random vector $\vec{y}$ by

$$\vec{y} = g(\vec{x})$$

↖ deterministic,
invertible,
continuous,
differentiable

In 1 dim:

$$p_y(y) = p_x(g^{-1}(y)) \left| \frac{\partial x}{\partial y} \right| \qquad p_x(x) = p_y(g(x)) \left| \frac{\partial g(x)}{\partial x} \right|$$

In n-dim:

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x})) \underbrace{\left| \det \left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}} \right) \right|}_{\text{Jacobian determinant}}$$

Unit 1: Background

1.2 Classical Statistics and Data Analysis Background

Sources:
- Cowan - Statistical data analysis
- also mostly covered in deeplearningbook.org

Estimators

- A (point) estimator makes a prediction for some quantity of interest λ.

  E.g. estimator of the mean height $\bar{h}$

  $$\hat{h} = \frac{1}{N} \sum_i h_i$$

  We write $\hat{\lambda}$, where „hat" means estimator.

- Given a dataset $\{x^{obs}\}$ drawn from / sampled a PDF $P(x)$, an estimator is some function

  $$\hat{\lambda} = \mathcal{E}[\{x^{obs}\}]$$

  some function, „guessed" or derived

- An optimal estimator is
  - unbiased : $\langle \hat{\lambda} \rangle = \lambda$ „correct on average"
  - minimum variance : $Var[\hat{\lambda}]$ is as small as possible (smallest error)

Some common estimators

notation: $\bar{x}$
„bar" means mean

(sample) mean: $\hat{\bar{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i^{obs}$

variance: $\hat{V} = \frac{1}{n-1} \sum_{n=1}^{n} (x_i - \hat{\bar{x}})^2$

covariance: $\widehat{Cov} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\bar{x}})(y_i - \hat{\bar{y}})$

These estimators have a variance, e.g.

$V[\hat{\bar{x}}] = \frac{\sigma^2}{n}$ where σ^2 is the variance of x

Likelihood, Posterior, Prior

- The Likelihood is the probability of measuring data $\vec{d}$ given a model M with parameters $\vec{\lambda}$.

$$\mathcal{L}(\vec{d} \mid \vec{\lambda}, M)$$

often not written out

The likelihood often appears as a loss function in machine learning.

It does not tell us the probability of $\vec{\lambda}$ given $\vec{d}$.

- This is given by the Posterior

$$P(\vec{\lambda}, M \mid \vec{d})$$

- Bayes theorem

$$P(\vec{\lambda} \mid d) = \frac{\mathcal{L}(\vec{d} \mid \vec{\lambda})\, P(\vec{\lambda})}{P(\vec{d})}$$

Course logistics

- **Reading for this lecture:**
  - **For example:** Deeplearningbook.org chapter 3, parts of chapter 5.

- **Problem set**: No problem set in the first week