TECHNISCHE UNIVERSITÄT DARMSTADT

Statistical Machine Learning

Lecture 03: Statistics Refresher

Kristian Kersting

TU Darmstadt

Summer Term 2020

Today's Objectives

- Make you remember your sweetest high school dreams: statistics & probabilities.
- This topic is harder than most of remaining chapters, but you will need it to continue!
- Covered Topics:
  - Random Variables: discrete & continuous
  - Distributions: discrete & continuous
- Expected values and moments
- Joint distributions, conditional distributions, independence

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Random Variables

- What is a random variable?
  - Is a random number determined by chance
  - More formally, drawn according to a probability distribution
  - Typical random variables in statistical learning: input data, output data, noise
- What is a probability distribution?
  - Describes the probability (density) that the random variable will be equal to a certain value.
  - The probability distribution can be given by the physics of an experiment (e.g., throwing dice)

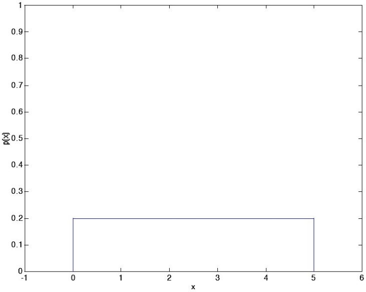
TECHNISCHE UNIVERSITÄT DARMSTADT

Random Variables

- Important concept: The data generating model
  - E.g., what is the data generating model for: i) throwing dice, ii) regression, iii) classification, iv) visual perception?
- Problem: On which time scale is a distribution observed?

TECHNISCHE UNIVERSITÄT DARMSTADT

Uniform Distribution



- All data is equally probable within a bounded region R

$$p(x) = \frac{1}{R}$$

- The uniform distribution plays an important role in entropy methods and information theory.

TECHNISCHE UNIVERSITÄT DARMSTADT

Discrete Distributions

- The random variables take on discrete values
  - E.g, when throwing a dice, the possible values are (countably finite set):

$$x_i \in \{1, 2, 3, 4, 5, 6\}$$

  - E.g., the number of sand grains at the beach (countably infinite set):

$$x_i \in \mathbb{N}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Discrete Distributions

- The probabilities sum to 1

$$\sum_i p(x_i) = 1$$

- Discrete distributions are particularly important in classification and decision making
- A discrete distribution is described by a probability mass function (or frequency function), which is a normalized histogram

TECHNISCHE UNIVERSITÄT DARMSTADT

Bernoulli Distribution

- A Bernoulli random variable only takes on two values, for example 0 and 1

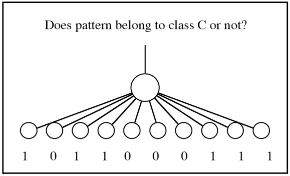
$$
\begin{aligned}
x &\in \{0, 1\} \\
p(x = 1|\mu) &= \mu \\
\text{Bern}(\text{x}|\mu) &= \mu^x (1 - \mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\text{var}[x] &= \mu(1 - \mu)
\end{aligned}
$$

- The only parameter of a Bernoulli distribution is μ, i.e., it is completely defined using only this parameter

TECHNISCHE UNIVERSITÄT DARMSTADT

Bernoulli Distribution

- Bernoulli distributions are often modeled with sigmoidal nonlinearites in statistical learning



TECHNISCHE UNIVERSITÄT DARMSTADT

Binomial Distribution

- Binomial variables are a sequence of N repeated Bernoulli variables
- One interpretation is "what is the probability of getting $m \in \mathbb{N}$ heads in N trials?"

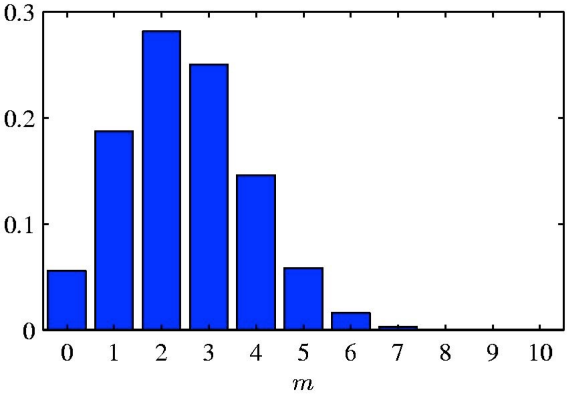
$$
\begin{aligned}
\text{Bin}(m|N,\mu) &= \begin{pmatrix} N \\ m \end{pmatrix} \mu^m (1-\mu)^{N-m} \\
\mathbb{E}[m] &= \sum_{m=0}^{N} m\text{Bin}(m|N,\mu) = N\mu \\
\text{var}[m] &= \sum_{m=0}^{N} (m-\mathbb{E}[m])^2 \text{Bin}(m|N,\mu) = N\mu(1-\mu)
\end{aligned}
$$

Binomial Distribution

- The Binomial distribution is completely defined with *N* - the number of samples - and μ- the probability that one sample is equal to 1
- Binomial variables are important for example in density estimation: “What is the probability that *k* out of *n* data points fall into region *R*?”

Binomial Distribution

$\mathrm{Bin}(m|10, 0.25)$



TECHNISCHE UNIVERSITÄT DARMSTADT

Multinoulli Distribution

- Multinoulli variables, also called Categorical variables in some literature, are a generalization of binomial variables to multiple outputs (e.g., multiple classes)
- 1-of-K coding scheme (also called one-hot encoding)

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^\intercal$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \quad \forall k : \mu_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

$$\mathbb{E}\left[\mathbf{x}|\boldsymbol{\mu}\right] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^\intercal$$

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} u_k = 1$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Multinomial Distribution

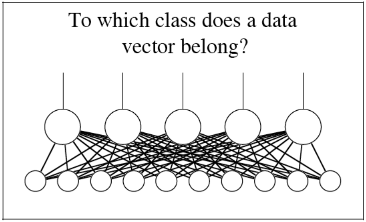
- N independent trials can result in one of K types of outcome
- What is the probability that in N trials, the frequency of the K classes is $m_1, m_2, \ldots, m_K$

$$
\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_k | \boldsymbol{\mu}, N) &= \binom{N}{m_1, m_2, \ldots, m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\
\mathrm{cov}\left[m_j m_k\right] &= -N\mu_j\mu_k
\end{aligned}
$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Multinomial Distribution

- The multinomial distribution play an important role in multi-class classification ($N = 1$)



TECHNISCHE UNIVERSITÄT DARMSTADT

Poisson Distribution

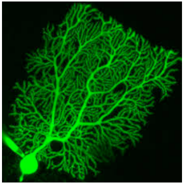
- The Poisson distribution is the binomial distribution where the number of trials N goes to infinity, and the probability of success on each trial, μ, goes to zero, such that $N\mu = \lambda$ is a constant

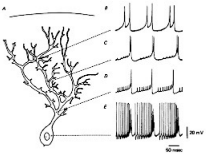
$$p(m|\lambda) = \frac{\lambda^m}{m!} e^{-\lambda}$$

- Where the m is the number of "successes"
- For example, Poisson distributions are an important model for t he firing characteristics of biological neurons. They are also used as an approximation to binomial variables with small p

Poisson Distribution

- Example: What is the probability of firing of a *Purkinje* neuron in the cerebellum in a 10ms time interval?
  - We know that the average firing of these neurons is about 40Hz, $\lambda = 40\text{Hz} \times 0.01\text{s}$
  - Note that this approximation only work if the number of spike is low in the given time interval





TECHNISCHE UNIVERSITÄT DARMSTADT

Continuous Distributions

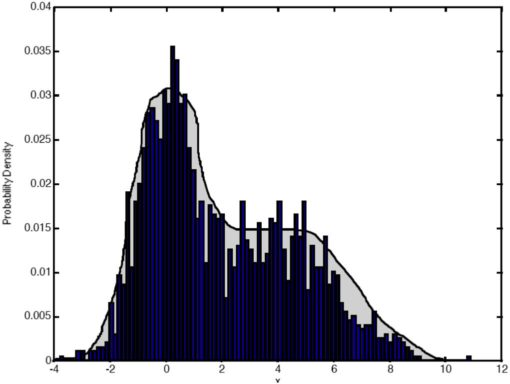
- The random variables take on continuous values
- Continuous distributions are discrete distributions where the number of discrete values goes to infinity, while the probability of each value goes to zero
- A continuous distribution is described by a probability density function, which integrates to 1

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

- Continuous distributions are particularly important in regression and unsupervised learning
- A lot of Machine Learning is centered around how to better model a density function

TECHNISCHE UNIVERSITÄT DARMSTADT

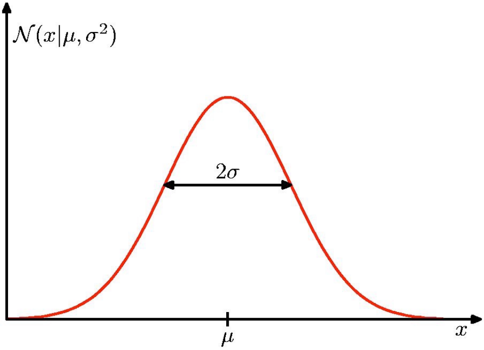
Example of a probability density function $p(x)$



$$P(a < x < b) = \int_a^b p(x)dx$$

TECHNISCHE UNIVERSITÄT DARMSTADT

The Gaussian Distribution



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

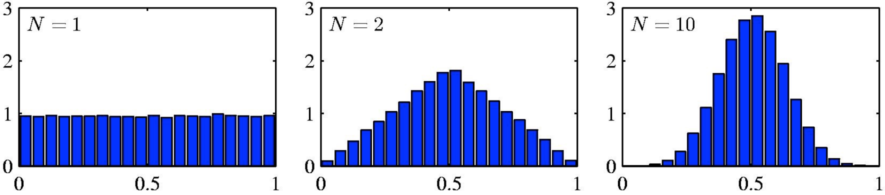
Central Limit Theorem

- Why are Gaussians SO important?
- The distribution of the sum of N i.i.d. (independent and identically distributed) random variables becomes increasingly Gaussian as N grows

TECHNISCHE UNIVERSITÄT DARMSTADT

Central Limit Theorem

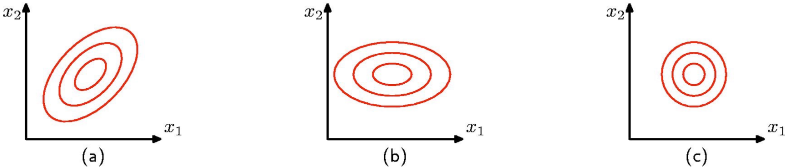
- Example: *N* uniform [0,1] random variables



- Gaussians are often a *good* model of data
- Working with Gaussians leads to analytic solutions for complex operations

TECHNISCHE UNIVERSITÄT DARMSTADT

The Multivariate Gaussian Distribution



$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

The Multivariate Gaussian Distribution

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- **To clear some confusion**: for a chosen vector $\mathbf{x}$, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a real number with the probability density of $\mathbf{x}$ (which can be greater than 1, only the integral of the probability density function needs to be 1). The mean $\boldsymbol{\mu}$ is just a specific vector amongst all the possible vectors. The covariance matrix $\boldsymbol{\Sigma}$ tells us how two dimensions of a vector are related to each other.

TECHNISCHE UNIVERSITÄT DARMSTADT

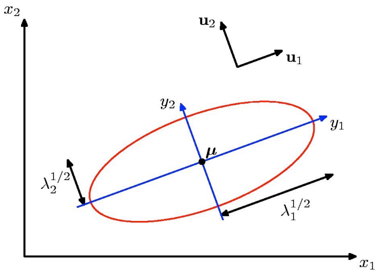
Geometry of the Multivariate Gaussian

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^\mathsf{T}$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^\mathsf{T} (\mathbf{x} - \boldsymbol{\mu})$$



Δ^2 is the Mahalanobis distance.

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

2. Basic Rules of Probability

- Joint Distribution

$$p(x, y)$$

- Marginal Distribution

$$p(y) = \int p(x, y)dx$$

- Conditional Distribution

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

2. Basic Rules of Probability

- Probabilistic Independence

$$p(x, y) = p(x)p(y)$$

- Chain Rule of Probabilities

$$\begin{aligned} p(x_1, \ldots, x_n) &= p(x_1|x_2, \ldots, x_n)p(x_2, \ldots, x_n) \\ &= p(x_1|x_2, \ldots, x_n)p(x_2|x_3, \ldots, x_n) \ldots p(x_{n-1}|x_n)p(x_n) \end{aligned}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Bayes Rule

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

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

- **posterior**: $p(y|x)$
- **likelihood:** $p(x|y)$
- **prior:** $p(y)$
- $p(x) = \int p(x,y)\mathrm{d}y = \int p(x|y)p(y)\mathrm{d}y$

Partitioned Gaussian Distributions

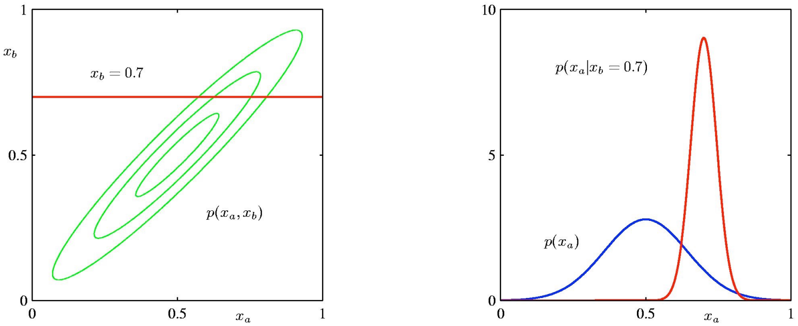
$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$\boldsymbol{\Lambda}$ is the precision matrix.

Partitioned Conditionals and Marginals



TECHNISCHE UNIVERSITÄT DARMSTADT

Partitioned Conditionals and Marginals

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b}\right)$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{a|b} &= \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba} \\
&= \boldsymbol{\Sigma}_{a|b}\left\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu} - \boldsymbol{\Lambda}_{ab}\left(\mathbf{x}_b - \boldsymbol{\mu}\right)\right\} \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\left(\mathbf{x}_b - \boldsymbol{\mu}_b\right)
\end{aligned}$$

$$\begin{aligned}
p\left(\mathbf{x}_a\right) &= \int p\left(\mathbf{x}_a, \mathbf{x}_b\right) d\mathbf{x}_b \\
&= \mathcal{N}\left(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}\right)
\end{aligned}$$

- Important result: If the joint distribution $p(\mathbf{x_a}, \mathbf{x_b})$ is Gaussian, then the conditional distributions $p(\mathbf{x_a}|\mathbf{x_b})$ and $p(\mathbf{x_b}|\mathbf{x_a})$ are also Gaussians. Moreover, the marginal distributions $p(\mathbf{x_a})$ and $p(\mathbf{x_b})$ are also Gaussians

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Expectations

- Expectation

$$\mathbb{E}_{x\sim p(x)}\left[f(x)\right] = \mathbb{E}_x\left[f\right] = \mathbb{E}\left[f\right] = \begin{cases} \sum_x p(x)f(x) & \text{discrete case} \\ \int p(x)f(x)\mathrm{d}x & \text{continuous case} \end{cases}$$

- Conditional Expectation

$$\mathbb{E}_{x\sim p(x|y)}\left[f\left(x\right)\right] = \mathbb{E}_x\left[f|y\right] = \begin{cases} \sum_x p(x|y)f(x) & \text{discrete case} \\ \int p(x|y)f(x)\mathrm{d}x & \text{continuous case} \end{cases}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Expectations

- Approximate Expectation

$$\mathbb{E}[f] = \int f(x)p(x)\mathrm{d}x \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

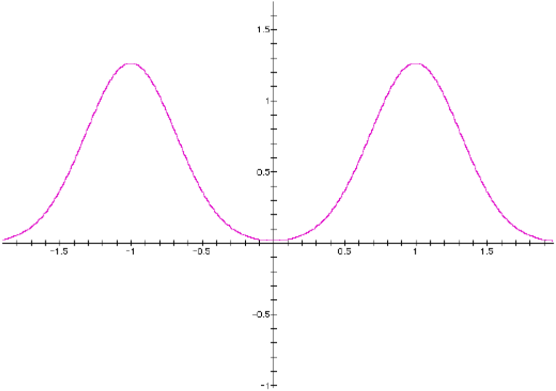
  - We sample N points from the distribution $p(x)$ and compute the function at those points. The probability of computing $f(x_n)$ for a certain point x_n is given by the probability of sampling $p(x_n)$

- This result is very important! When there is no analytical solution, we can use this to approximate integrals by sampling!

TECHNISCHE UNIVERSITÄT DARMSTADT

Expectations

- Example: What is the expectation of the following distribution?



TECHNISCHE UNIVERSITÄT DARMSTADT

Expectations

- Some rules of expectation

  $\mathbb{E}[a\mathbf{x}] = a\mathbb{E}[\mathbf{x}]$

  $\mathbb{E}[\mathbf{x}+\mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{y}]$

  $\mathbb{E}[\mathbf{x}\mathbf{y}] = \mathbb{E}[\mathbf{x}]\,\mathbb{E}[\mathbf{y}]$ only if **x** and **y** are statistically independent!

  $\mathbb{E}\left[\sum_i a_i x_i\right] = \sum_i a_i \mathbb{E}[x_i]$

- Expectation of functions

  $\mathbb{E}[g(\mathbf{x})] = \int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$

  In general $\mathbb{E}[g(\mathbf{x})] \neq g(\mathbb{E}[\mathbf{x}])$

TECHNISCHE UNIVERSITÄT DARMSTADT

Variance and Covariance

- Variances give a measure of dispersion - the expected spread of the variable in relation to its mean

$$\mathrm{var}[x] = \mathbb{E}\left[(x - \mathbb{E}[x])^2\right] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Variance and Covariance

- Covariances give a measure of correlation - how much two variables change together

$$\begin{aligned}\operatorname{cov}[x,y] &= \mathbb{E}_{x,y}\left[(x-\mathbb{E}[x])(y-\mathbb{E}[y])\right] \\ &= \mathbb{E}_{x,y}[xy]-\mathbb{E}_x[x]\mathbb{E}_y[y]\end{aligned}$$

$$\begin{aligned}\operatorname{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{y}-\mathbb{E}[\mathbf{y}])^\intercal\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{y}^\intercal-\mathbb{E}[\mathbf{y}^\intercal])\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^\intercal]-\mathbb{E}_{\mathbf{x}}[\mathbf{x}]\mathbb{E}_{\mathbf{y}}[\mathbf{y}^\intercal]\end{aligned}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Variance and Covariance

- Note the very important rule

$$\begin{aligned}\mathbb{E}\left[\mathbf{x}\mathbf{x}^\intercal\right] &= \mathbb{E}_{\mathbf{x}}[\mathbf{x}]\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}^\intercal\right] + \operatorname{cov}\left[\mathbf{x}, \mathbf{x}\right] \\ &= \boldsymbol{\mu}\boldsymbol{\mu}^\intercal + \boldsymbol{\Sigma}\end{aligned}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Moments of Random Variables

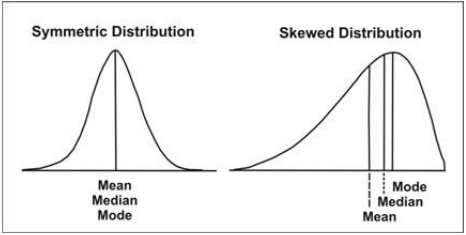
- Definition of a Moment

  $m_n = \mathbb{E}\left[x^n\right]$

- Definition of a Central Moment

  $cm_n = \mathbb{E}\left[(x - \mu)^n\right]$

- cm_2: variance
- cm_3: skewness (measure of asymmetry)
- cm_4: kurtosis (measure of heavy tailed-ness and light tailed-ness)



Moments of the Multivariate Gaussian

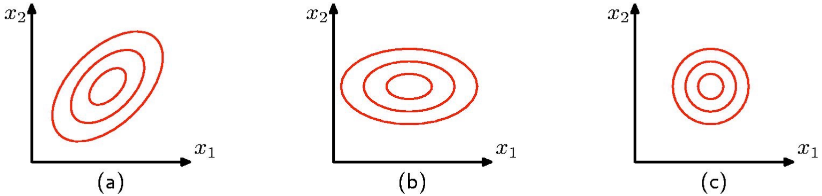
$$\mathbb{E}[\mathbf{x}] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\} \mathbf{x} d\mathbf{x}$$
$$= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{z}\right\} (\mathbf{z}+\boldsymbol{\mu})\, d\mathbf{z}$$

Thanks to the asymmetry of **z**, $\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$

Moments of the Multivariate Gaussian

$$\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\intercal}\right] = \boldsymbol{\mu}\boldsymbol{\mu}^{\intercal} + \boldsymbol{\Sigma}$$

$$\operatorname{cov}\left[\mathbf{x}\right] = \operatorname{cov}\left[\mathbf{x}, \mathbf{x}\right] = \mathbb{E}\left[\left(\mathbf{x} - \mathbb{E}\left[\mathbf{x}\right]\right)\left(\mathbf{x} - \mathbb{E}\left[\mathbf{x}\right]\right)^{\intercal}\right] = \boldsymbol{\Sigma}$$



TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

4. Exponential Family

- The exponential family are a large class of distributions that are all analytically appealing, because taking the log of them decomposes them into simple terms
- All distributions from this family are uni-modal

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})\, g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\}$$

where $\boldsymbol{\eta}$ is the natural parameter and

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\{\boldsymbol{\eta}^\mathsf{T}\mathbf{u}(\mathbf{x})\}\, \mathrm{d}\mathbf{x} = 1$$

hence g can be interpreted as a normalization coefficient

TECHNISCHE UNIVERSITÄT DARMSTADT

Exponential Family - Bernoulli Distribution

- The Bernoulli Distribution

$$\begin{aligned} p(x|\mu) = \mathrm{Bern}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\ &= \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned}$$

- Comparing with the general form we see that

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right), \quad \mu = \underbrace{\sigma(\eta) = \frac{1}{1+\exp(-\eta)}}_{\text{Logistic sigmoid}}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Exponential Family - Bernoulli Distribution

- Hence, the Bernoulli Distribution can be written as

$$p\left(x|\mu\right) = \sigma(-\eta)\exp(\eta x)$$

where

$$u(x) = x, \quad h(x) = 1, \quad g\left(\eta\right) = 1 - \sigma(\eta) = \sigma(-\eta)$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Exponential Family - Multinoulli Distribution

- The Multinoulli Distribution also belongs to the exponential family

$$p\left(\mathbf{x}|\boldsymbol{\mu}\right) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\right\}$$

where

$$\mathbf{x} = \left(x_1, \ldots, x_M\right)^\intercal, \quad \boldsymbol{\eta} = \left(\eta_1, \ldots, \eta_M\right)^\intercal, \quad \eta_k = \ln u_k$$
$$\mathbf{u}(\mathbf{x}) = \mathbf{x}, \quad h(\mathbf{x}) = 1, \quad g(\boldsymbol{\eta}) = 1$$

- Note that the parameters η_k have to be chosen in a way to guarantee that $p\left(\mathbf{x}|\boldsymbol{\mu}\right)$ is a valid probability distribution. Particularly, they must satisfy

$$\sum_{\mathbf{x}} p\left(\mathbf{x}|\boldsymbol{\mu}\right) = 1 \Longrightarrow \sum_{k=1}^{M} \mu_k = 1$$

Exponential Family - Multinoulli Distribution

- Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$, which ensures that the distribution is well defined. We can rewrite $p(\mathbf{x}|\boldsymbol{\mu})$ and observe that

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right), \quad \mu_k = \underbrace{\frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}}_{\text{Softmax}}$$

- Here the parameters η_k can be chosen independently, since

$$0 \leq \mu_k \leq 1, \quad \sum_{k=1}^{M-1} \mu_k \leq 1$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Exponential Family - Multinoulli Distribution

- The Multinoulli Distribution can then be written as

$$p(\mathbf{x}|\boldsymbol{\mu}) = h(\mathbf{x})\, g(\boldsymbol{\eta}) \exp\{\boldsymbol{\eta}^\intercal \mathbf{u}(\mathbf{x})\}$$

where

$$\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{M-1}, 0)^\intercal, \quad \mathbf{u}(\mathbf{x}) = \mathbf{x}, \quad h(\mathbf{x}) = 1$$

$$g(\boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Exponential Family - Gaussian Distribution

- The Gaussian Distribution can be rewritten as

$$
\begin{aligned}
p(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \\
&= h(x)g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^\intercal \mathbf{u}(x)\right\}
\end{aligned}
$$

where

$$
\boldsymbol{\eta} = \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right)^\intercal, \quad \mathbf{u}(x) = \left(x^2, x\right)^\intercal, \quad h(\mathbf{x}) = 1
$$

$$
g(\boldsymbol{\eta}) = \sqrt{\frac{-\eta_1}{\pi}} \exp\left(\frac{\eta_2^2}{4\eta_1}\right)
$$

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Information Theory - Core Questions

- Classical Question: How can we represent information compactly, i.e., using as few bits as possible?
  - Compressing text like with GZIP
  - Compressing pictures like in JPEG, movies like in MPEG
  - Compressing sound using MP3
- Classical Question: How can we transmit or store data reliably?
  - ECC memory
  - Error Correction on CDs
  - Communication with space probes

Information Theory - Core Questions

- Machine Learning Questions:
  - How can we measure complexity?
  - How can we measure "distances" between probability distributions?
  - How can we reconstruct data?
- We are not covering all questions here... :)

TECHNISCHE UNIVERSITÄT DARMSTADT

What is Information?

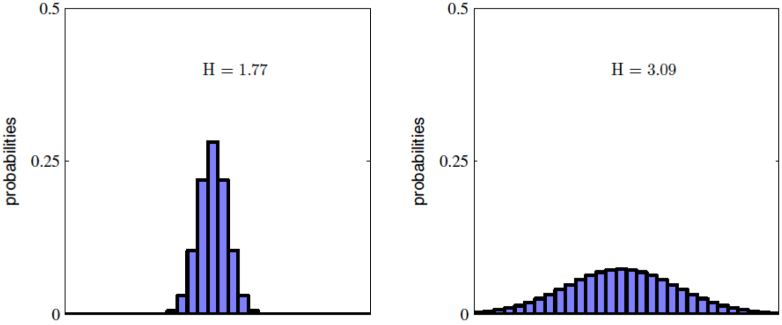
i	a_i	p_i
1	a	.0575
2	b	.0128
3	c	.0263
4	d	.0285
5	e	.0913
6	f	.0173
7	g	.0133
8	h	.0313
9	i	.0599
10	j	.0006
11	k	.0084
12	l	.0335
13	m	.0235
14	n	.0596
15	o	.0689
16	p	.0192
17	q	.0008
18	r	.0508
19	s	.0567
20	t	.0706
21	u	.0334
22	v	.0069
23	w	.0119
24	x	.0073
25	y	.0164
26	z	.0007
27	-	.1928

- All letters in the English alphabet have a very different probability p_i of occurring
- What is the number of bits you need to represent 27 characters? $\lceil \log_2 27 \rceil \approx \lceil 4.75 \rceil = 5$ bits
- How can we measure the information in a single character? $h(p_i) = -\log_2 p_i$. Events with a low probability correspond to high information content
- So, what is the average information in a character in an English text?
  - $H(p) = \mathbb{E}[h(.)] = \sum_i p_i h(p_i) = -\sum_i p_i \log_2 p_i \approx 4.1$

  This quantity is called the entropy. On average, with the right encoding, we can represent each letter with 4.1 bits instead of 4.7

TECHNISCHE UNIVERSITÄT DARMSTADT

Entropy of Distributions



What is the "difference" between these distributions?

TECHNISCHE UNIVERSITÄT DARMSTADT

Kullback-Leibler Divergence

- The Kullback-Leibler Divergence - KL Divergence - is a similarity measure between two distributions, and is defined as

$$\begin{aligned}\mathrm{KL}\left(p\|q\right) &= -\int p(x)\ln q(x)\mathrm{d}x - \left(-\int p(x)\ln p(x)\mathrm{d}x\right)\\ &= -\int p(x)\ln\frac{q(x)}{p(x)}\mathrm{d}x\end{aligned}$$

- It represents the average additional amount of extra bits required to specify a symbol x, given that its underlying probability distribution is the estimated $q(x)$ and not the true one $p(x)$

Kullback-Leibler Divergence

- Some properties
  - It is not a distance: $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$
  - It is non-negative: $\mathrm{KL}(p||q) \geq 0$
  - If $\forall x\ p(x) = q(x)$: $\mathrm{KL}(p||q) = 0$
- There are other metrics of similarity, but as we will see further in the course, the KL Divergence is deeply connected with maximum likelihood estimation

TECHNISCHE UNIVERSITÄT DARMSTADT

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

6. Wrap-Up

You know now:

- What random variables are (both continuous and discrete)
- What probability distributions are
- Some basic rules of probability theory
- What expectation and variance are
- What a Gaussian distribution is and why it is so important
- What information and entropy are
- How to measure the similarity between two probability distributions

TECHNISCHE UNIVERSITÄT DARMSTADT

Self-Test Questions

- What is a random variable?
- What is a distribution?
- What is a Binomial distribution?
- How does a Poisson distribution relate to Binomial distributions?
- What is a Gaussian distribution?
- What is an expectation?
- What is a joint distribution?
- What is a conditional distribution?
- What is a distribution with a lot of information?
- How to measure the difference between distributions?

Homework

- Reading Assignment for next lecture
  - Bishop appendix E