Computational Statistics
Primer on probability and statistics

Luca Bortolussi

Department of Mathematics and Geosciences
University of Trieste

Office 303, third floor, E1 3
`luca@dmi.units.it`

Trieste, Winter Semester 2015/2016

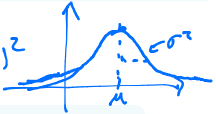
Outline

1. Basics of probability theory
2. Some probability distributions
3. Fitting distributions
4. Information theory
5. Decision theory

MULTIVARIATE NORMAL DISTRIBUTION

1D

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



$\mu_i = \langle x_i \rangle = E[x_i]$

$\Sigma_{ij} = cov(x_i, x_j)$

This is the most important distribution we will use, and generalises the 1d normal. In *d* dimensions

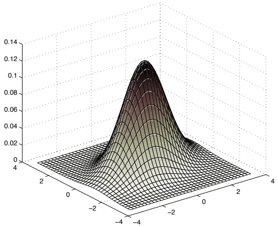
d-dim vector

Σ symm + pos. defined

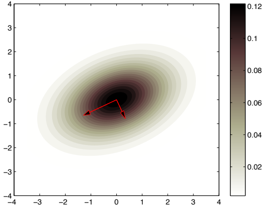
$$p(\mathbf{x}|\underbrace{\boldsymbol{\mu}}, \underbrace{\boldsymbol{\Sigma}}_{d\times d}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\right)$$

quadratic form

It holds $\boldsymbol{\mu} = \langle \mathbf{x} \rangle$, and $\boldsymbol{\Sigma} = \mathrm{cov}(\mathbf{x}, \mathbf{x}) = \langle (\mathbf{x}-\boldsymbol{\mu})\overset{T}{(}\mathbf{x}-\boldsymbol{\mu})^T \rangle$



(a)



(b)

PROPERTIES OF MULTIVARIATE NORMAL

Completing the square

A useful technique in manipulating Gaussians is completing the square. For example, the expression

$$\exp\left(-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}+\mathbf{b}^\mathsf{T}\mathbf{x}\right) \qquad (8.4.10)$$

can be transformed as follows. First we complete the square:

$$\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}-\mathbf{b}^\mathsf{T}\mathbf{x}=\frac{1}{2}\left(\mathbf{x}-\mathbf{A}^{-1}\mathbf{b}\right)^\mathsf{T}\mathbf{A}\left(\mathbf{x}-\mathbf{A}^{-1}\mathbf{b}\right)-\frac{1}{2}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b} \qquad (8.4.11)$$

Hence

$$\exp\left(-\frac{1}{2}\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}+\mathbf{b}^\mathsf{T}\mathbf{x}\right)=\mathcal{N}\left(\mathbf{x}|\underline{\mathbf{A}^{-1}\mathbf{b}},\underline{\mathbf{A}^{-1}}\right)\sqrt{\det\left(2\pi\mathbf{A}^{-1}\right)}\exp\left(\frac{1}{2}\mathbf{b}^\mathsf{T}\mathbf{A}^{-1}\mathbf{b}\right) \qquad (8.4.12)$$

~~$p(\mathbf{x}|\mathbf{A},\mathbf{b})$~~ $c\exp\left(-\frac{1}{2}\mathbf{x}\mathbf{A}\mathbf{x}^T+\mathbf{b}^T\mathbf{x}\right)$ is known as the canonical representation, and it is normal with mean $\boxed{\mathbf{A}^{-1}\mathbf{b}}$ and covariance $\boxed{\mathbf{A}^{-1}.}$

Linear transformation

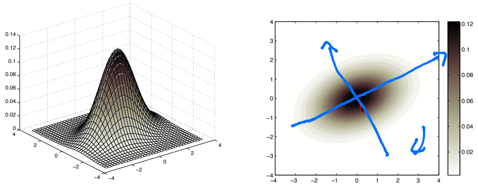
Result 8.3 (Linear Transform of a Gaussian)**.** Let $\mathbf{y}$ be linearly related to $\mathbf{x}$ through

$$\boxed{\mathbf{y}}=\mathbf{M}\underline{\mathbf{x}}+\underline{\boldsymbol{\eta}} \qquad (8.4.14)$$

where $\mathbf{x}\perp\!\!\!\perp\boldsymbol{\eta}$, $\boldsymbol{\eta}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, and $\mathbf{x}\sim\mathcal{N}(\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)$. Then the marginal $p(\mathbf{y})=\int_\mathbf{x}p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ is a Gaussian

$$\rightsquigarrow\ p(\mathbf{y})=\mathcal{N}\left(\mathbf{y}|\underline{\mathbf{M}\boldsymbol{\mu}_x+\boldsymbol{\mu}},\underline{\mathbf{M}\boldsymbol{\Sigma}_x\mathbf{M}^\mathsf{T}+\boldsymbol{\Sigma}}\right) \qquad (8.4.15)$$

PROPERTIES OF MULTIVARIATE NORMAL



(a)

(b)

Eigendecomposition

$$\mathbf{\Sigma} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\mathsf{T}} \tag{8.4.5}$$

where $\mathbf{E}^{\mathsf{T}}\mathbf{E} = \mathbf{I}$ and $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_D)$. In the case of a covariance matrix, all the eigenvalues λ_i are positive. This means that one can use the transformation

$$\mathbf{y} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{E}^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}) \tag{8.4.6}$$

so that

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}\mathbf{E}\mathbf{\Lambda}^{-1}\mathbf{E}^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}^{\mathsf{T}}\mathbf{y} \tag{8.4.7}$$

So by rescaling, we can obtain a product of *d*-univariate standard normal distributions, one per dimension.

PROPERTIES OF MULTIVARIATE NORMAL

Marginal and conditional of multivariate Gaussians

Result 8.4 (Partitioned Gaussian)**.** Consider a distribution $\mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ defined jointly over two vectors $\mathbf{x}$ and $\mathbf{y}$ of potentially differing dimensions,

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \Delta\sim \tag{8.4.16}$$

with corresponding mean and partitioned covariance

$$\boldsymbol{\mu} = \left(\begin{array}{c} \boldsymbol{\mu}_x \\ \hline \boldsymbol{\mu}_y \end{array}\right) \qquad \boldsymbol{\Sigma} = \left(\begin{array}{c|c} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \hline \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{array}\right) \tag{8.4.17}$$

where $\boldsymbol{\Sigma}_{yx} \equiv \boldsymbol{\Sigma}_{xy}^{\mathsf{T}}$. The marginal distribution is given by

$$\rightsquigarrow \quad p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}) \qquad p(x) = \int p(x,y)\,dy \tag{8.4.18}$$

and conditional

$$\rightsquigarrow \quad p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y), \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right) \tag{8.4.19}$$

PROPERTIES OF MULTIVARIATE NORMAL

Product of multivariate Gaussians

Result 8.2 (Product of two Gaussians)**.** The product of two Gaussians is another Gaussian, with a multiplicative factor, exercise(8.35):

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\,\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\,\frac{\exp\left(-\frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\mathbf{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)}{\sqrt{\det(2\pi\mathbf{S})}} \tag{8.4.8}$$

where $\mathbf{S} \equiv \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2$ and the mean and covariance are given by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}_1\mathbf{S}^{-1}\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_2\mathbf{S}^{-1}\boldsymbol{\mu}_1 \qquad \boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1\mathbf{S}^{-1}\boldsymbol{\Sigma}_2 \tag{8.4.9}$$

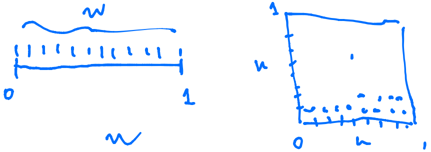
Gaussian average of a quadratic function

Result 8.5 (Gaussian average of a quadratic function)**.**

$$\left\langle \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} \right\rangle_{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})} = \boldsymbol{\mu}^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu} + \mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma})$$

THE CURSE OF DIMENSIONALITY

Exercise I: Suppose you want to explore uniformly a region by gridding it. How many grid points do you need?



n^2

↑

crescita esponenziale in d

d? $\underline{n^d}$

THE CURSE OF DIMENSIONALITY

Exercise II: Suppose you sample from a uniform distribution in d dimensions. What is the probability of finding a point inside the region $[\epsilon, 1-\epsilon]^d$?

$U = (U_1, \ , U_d)$ $\quad U_i \sim \text{UNIFORM}(0,1)$

$$\vec{u} \in \underbrace{[\epsilon, 1-\epsilon]^d}_{R_\epsilon} \iff u_i \in [\epsilon, 1-\epsilon] \qquad (P(u_i)=1)$$

$$p(u_i \in [\epsilon, 1-\epsilon]) = 1-2\epsilon$$

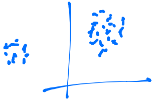
$$\forall \epsilon, \ P(\vec{u} \in R_\epsilon) = (1-2\epsilon)^d \xrightarrow[d \to \infty]{} 0$$

THE CURSE OF DIMENSIONALITY

d-dim, indip. standard gauss,

Exercise III: Suppose you sample from a spherical Gaussian distribution. Where do the points lie as the dimensions increase?

MIXTURES: HOW TO BUILD MORE DISTRIBUTIONS



- More general distributions can be built via mixtures: e.g.

$$p(x|\mu_{1,\dots,n},\sigma^2_{1,\dots,n}) = \sum_i \underset{[0,1]}{\boxed{\pi_i}}\, \underbrace{\mathcal{N}(\mu_i,\sigma_i^2)}$$

where the *mixing coefficients* π_i are discretely distributed , $\sum \pi_i = 1$

MIXTURES: HOW TO BUILD MORE DISTRIBUTIONS

- More general distributions can be built via mixtures: e.g.

$$p(x|\mu_{1\dots,n},\sigma^2_{1,\dots,n}) = \sum_i \pi_i \mathcal{N}(\mu_i,\sigma_i^2)$$

  where the *mixing coefficients* π_i are discretely distributed
- You can interpret this as a two stage hierarchical process: choose one component out of a discrete distribution, then choose the distribution for that component

MIXTURES: HOW TO BUILD MORE DISTRIBUTIONS

- **IMPORTANT CONCEPT**: the mixture

$$p(x|\mu_{1,\dots,n},\sigma^2_{1,\dots,n}) = \sum_i \underbrace{\pi_i}\underbrace{\mathcal{N}(\mu_i,\sigma_i^2)}$$

is an example of *latent variable model*, with a latent class variable and an observed continuous value. The mixture is the marginal distribution for the observations (w.r.t. the latent variable)

$$P(x) = \sum_{\pi} p(x,\pi) = \sum_{\pi} P(\pi)\cdot p(x|\pi)$$

$$p(\pi) = \sum_{x} p(\pi,x) = \sum_{x} p\underline{(\pi|x)}\,p(x)$$

$\vec{x}$ oss. $p(\pi|\vec{x})$ ← Bayes.

MIXTURES: HOW TO BUILD MORE DISTRIBUTIONS

- **IMPORTANT CONCEPT**: the mixture

$$p(x|\mu_{1\dots,n},\sigma^2_{1,\dots,n}) = \sum_i \pi_i \mathcal{N}(\mu_i,\sigma_i^2)$$

  is an example of *latent variable model*, with a latent class variable and an observed continuous value. The mixture is the marginal distribution for the observations (w.r.t. the latent variable)

- The probability of the latent variables given the observations can be obtained using Bayes' theorem.

Continuous mixtures: some cool distributions

- No need for the mixing distribution (latent variable) to be discrete

CONTINUOUS MIXTURES: SOME COOL DISTRIBUTIONS

- No need for the mixing distribution (latent variable) to be discrete
- Suppose you are interested in the means of normally distributed samples (possibly with different variances/ precisions): Marginalising the precision in a Gaussian using a Gamma mixing distribution yields a *Student t-distribution*

CONTINUOUS MIXTURES: SOME COOL DISTRIBUTIONS

- No need for the mixing distribution (latent variable) to be discrete
- Suppose you are interested in the means of normally distributed samples (possibly with different variances/ precisions): Marginalising the precision in a Gaussian using a Gamma mixing distribution yields a *Student t-distribution*
- Suppose you have multiple rare event processes happening with slightly different rates: Marginalising the rate in a Poisson distribution using a Gamma mixing distribution yields a *negative binomial* distribution

Outline

1. Basics of probability theory
2. Some probability distributions
3. Fitting distributions
4. Information theory
5. Decision theory

PARAMETERS?

- Many distributions are written as conditional probabilities *given* the parameters: $p(x|\theta)$
- Often the values of the parameters are not known
- If we have observations, we can try to estimate the parameters from such data.
- We assume to have independent and identically distributed (i.i.d.) observations of $p(x|\theta_{true})$: $\mathbf{x} = x_1, \ldots, x_N$.

MAXIMUM LIKELIHOOD

- Likelihood for $\boxed{\text{i.i.d.}}$ observations $\mathbf{x} = x_1, \ldots, x_N$:

$$p(\mathbf{x}|\theta) \underset{\uparrow\ \text{independence}}{=} \boxed{\prod_{i=1}^{N} p(x_i|\theta)}$$

MAXIMUM LIKELIHOOD

- Likelihood for i.i.d. observations $\mathbf{x} = x_1, \ldots, x_N$:

$$p(\mathbf{x}|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

- Choose the parameters that best explain the observations: we pick θ by maximum likelihood:

$$\hat{\theta} = \mathrm{argmax}_\theta \left[\prod_i p(x_i|\theta) \right]$$

MAXIMUM A POSTERIORI

- Suppose we can encode prior knowledge (or absence of it) in a prior distribution over parameters, $p(\theta)$.
- We can then compute the posterior distribution, given i.i.d. observations $\mathbf{x} = x_1, \ldots, x_N$, by Bayes theorem:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

EVIDENCE
=
MARGINAL
LIKELIHOOD

where

$$\boxed{p(\mathbf{x}) = \int_\theta p(\mathbf{x}|\theta)p(\theta)d\theta}$$

MAXIMUM A POSTERIORI

- Suppose we can encode prior knowledge (or absence of it) in a prior distribution over parameters, $p(\theta)$.
- We can then compute the posterior distribution, given i.i.d. observations $\mathbf{x} = x_1, \ldots, x_N$, by Bayes theorem:

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$$

where

$$p(\mathbf{x}) = \int_{\theta} p(\mathbf{x}|\theta)p(\theta)d\theta$$

- Estimate θ_{true} by the *maximum a posteriori (MAP)* estimate

$$\hat{\theta}_{MAP} = \mathrm{argmax}_{\theta}\left[p(\theta)\prod_i p(x_i|\theta)\right]$$

EXERCISE: FITTING A DISCRETE DISTRIBUTION

- We have a discrete distribution with values in $K = \{1, \dots, k\}$, with parameters $\boldsymbol{\mu} = \mu_1, \dots, \mu_k$, $\sum_i \mu_i = 1$.
- We have independent observations $\mathrm{x} = x_1, \dots, x_N$, each taking values in K.
- The likelihood is

$$\mathcal{L}(\boldsymbol{\mu}) = p(\mathrm{x}|\boldsymbol{\mu}) = \prod_{i=1}^{N} p\,(x_i|\boldsymbol{\mu})$$

- Compute the Maximum Likelihood estimate of $\boldsymbol{\mu}$. What is the intuitive meaning of the result? What happens if one of the ~~D~~ K values is not represented in your sample?

EXERCISE: FITTING A DISCRETE DISTRIBUTION

$x_i = (0\cdots1\cdots0) = (x_{i1}, \ldots x_{iK})$ $\qquad \sum_j x_{ij} = 1$

$p(x_i|\vec{\mu}) = \prod_j \mu_j^{x_{ij}}$ (se $x_i \equiv J$, allora $p(x_i|\vec{\mu}) = \mu_J$)

$p(\vec{x}|\vec{\mu}) = \prod_{i,j} \mu_j^{x_{ij}} = \prod_j \mu_j^{n_j}$ $\qquad n_j = \#(j, \vec{x})$

$f(\vec{\mu}) = \log p(\vec{x}|\vec{\mu}) = \sum_j n_j \log \mu_j$

Impongo $\sum \mu_j = 1$

$\Downarrow$

λ - molt. di Lagr.

$f(\vec{\mu}, \lambda) = \sum_j n_j \log \mu_j + \lambda\left(\sum_j \mu_j - 1\right)$

$$\frac{\partial f}{\partial \mu_j} = \frac{n_j}{\mu_j} + \lambda = 0 \qquad \frac{\partial f}{\partial \lambda} = \sum_j \mu_j - 1 = 0$$

$$\mu_j = -\frac{n_j}{\lambda}$$

$$\lambda = -N$$

$$\boxed{\hat{\mu}_j = \frac{n_j}{N}}$$

Exercise II: fitting a Gaussian distribution

We have independent, real valued observations $\mathrm{x} = x_1, \ldots, x_N$.
Fit a Gaussian by maximum likelihood.

$$\prod_i p(x_i \mid \mu, \sigma^2) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$\mathcal{L}(\mu, \sigma^2) = -\frac{N}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \qquad \cdot \qquad \sum_i (x_i - \mu) = 0 \qquad \hat{\mu} = \frac{\sum x_i}{N}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = 0 \qquad -\frac{N}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2} \sum_i (x_i - \mu)^2 \cdot \frac{1}{(\sigma^2)^2} = 0 \qquad \hat{\sigma}^2 = \frac{\sum_i (x_i - \hat{\mu})^2}{N}$$

BAYESIAN ESTIMATION

- The Bayesian approach fully quantifies uncertainty
- The parameters are treated as additional random variables with their own *prior* distribution $p(\theta)$
- The observation likelihood is combined with the prior to obtain a *posterior* distribution via Bayes' theorem

$$p(\theta|\mathrm{x}) = \frac{p(\mathrm{x}|\theta)p(\theta)}{p(\mathrm{x})}$$

sum Σ

$$P(x) = \int P(x|\theta)P(\theta)d\theta$$

- The distribution of the observable x (*predictive* distribution) is obtained as

$$p(x|\mathrm{x}) = \int p(x|\theta)p(\theta|\mathrm{x})d\theta$$

EXERCISE: BAYESIAN FITTING OF GAUSSIANS

- Let data $x_i \quad i = 1, \ldots, N$ be distributed according to a Gaussian with mean μ and variance σ^2
- Let the prior distribution over the mean μ be a Gaussian with mean m and variance v^2
- Compute the posterior (and predictive distribution, Exercise) ↑ GAUSSIANA $\mathcal{N}(x \mid \mu_P, \sigma_P^2)$

$$\mu_P = m \cdot \frac{\sigma^2}{Nv^2 + \sigma^2} + \hat{\mu} \cdot \frac{Nv^2}{Nv^2 + \sigma^2}$$

$$\sigma_P^2 = \left(\frac{1}{v^2} + \frac{N}{\sigma^2} \right)^{-1}$$

EXERCISE: BAYESIAN FITTING OF GAUSSIANS

ESTIMATORS

- A procedure to calculate an expectation is called an *estimator*
- e.g., fitting a Gaussian to data by maximum likelihood provides the M.L. estimator for mean and variance, or Bayesian posterior mean
- An estimator will be a noisy estimate of the true value, due to finite sample effects
- An estimator $\hat{f}$ is *unbiased* if its expectation (under the joint distribution of the data set) coincides with the true value
- An estimator $\hat{f}$ is *consistent* if it converges to the true value when the number of data goes to infinity.

EXERCISE: BIASED ESTIMATOR

The ML estimator of variance, $\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$ is biased: $\langle\hat{\sigma}^2\rangle = \frac{N-1}{N}\underline{\sigma}^2$.

$\uparrow$ $P(\vec{x})$

$(\sigma^2 = \langle x_i^2\rangle - \mu^2$

$$\left[\begin{array}{l} \langle x_i\rangle = \mu \\ \langle x_i^2\rangle \; \sigma^2 + \mu^2 \\ \langle x_i x_j\rangle = \mu^2 \end{array}\right]$$

BOOTSTRAP

- For an estimator, in theory we can compute its mean and its variance under the joint distribution of the datasets. In practice, getting the variance may be very hard. Bootstrapping can be used instead.
- Given the dataset $\mathtt{x} = x_1, \ldots, x_N$, construct from it K new datasets $\mathtt{x}_\mathtt{i}$, also of size N, by sampling with repetitions.
- compute the estimator $\hat{\theta}_i$ for each $\mathtt{x}_\mathtt{i}$.
- Compute the empirical variance (or any other statistics) from $\mathtt{x}_1, \ldots, \mathtt{x}_\mathtt{K}$.
- This is an estimate of the actual variance of the estimator.

CONJUGATE PRIORS

- The Bayesian way has advantages in that it quantifies uncertainty and regularizes naturally
- BUT computing the normalisation in Bayes theorem is very hard
- The case when it is possible is when the prior and the posterior are of the same form (*conjugate*)
- Example + Exercise: Bernoulli and Beta.
- Example: discrete and Dirichlet
- Exercise: conjugate priors for the univariate normal (mean)

CONJUGATE PRIORS: ~~BINOMIAL~~ BERNOULLI AND BETA

Show that the Beta is the conjugate prior for the Bernoulli distribution.

BERNOULLI $x \in \{0,1\}$ $p(x|\vartheta) = \vartheta^{x}(1-\vartheta)^{1-x}$

$x_1, \dots, x_N$ $p(\vec{x}|\vartheta) = \prod_i p(x_i|\vartheta) = \vartheta^{K}(1-\vartheta)^{N-K}$

$K = \sum_i x_i = \#(1, \vec{x})$

$$\mathrm{Beta}(\vartheta|a,b) = c \cdot \vartheta^{a-1}(1-\vartheta)^{b-1}$$

$$p(\vartheta|\vec{x}) = \frac{1}{\boxed{p(\vec{x})}} \cdot \vartheta^{K}(1-\vartheta)^{N-K} \cdot \boxed{c}\,\vartheta^{a-1}(1-\vartheta)^{b-1}$$

$$= c'\,\vartheta^{a+K-1}(1-\vartheta)^{b+N-K-1} = \mathrm{Beta}(\vartheta\,|\,a+K-1, b+N-K-1)$$

Outline

1. Basics of probability theory
2. Some probability distributions
3. Fitting distributions
4. Information theory
5. Decision theory

ENTROPY

- Probability theory is the basis of information theory (interesting, but not the topic of this course).
- An important quantity is the *entropy* of a distribution

ENTROPY

- Probability theory is the basis of information theory (interesting, but not the topic of this course).
- An important quantity is the *entropy* of a distribution

$$H[p] = -\sum_i p_i \boxed{\log_2 p_i}$$

Or for continuous distributions:

$$\boxed{H[p] = -\int p(x) \log p(x) dx}$$

- Entropy measures the level of disorder of a distribution; for discrete distributions, it is always ≥ 0 and 0 only for deterministic distributions. The maximum is $\log K$, if K is the size of the support of the discrete distribution, and is achieved by the uniform distribution.

DIVERGENCE

- The *relative entropy* or *Kullback-Leibler (KL) divergence* between two distributions is

$$KL[\overset{\downarrow}{q}||\overset{\downarrow}{p}] = \sum_i q_i \log \frac{q_i}{p_i} = \underbrace{\langle \log q \rangle_q}_{H[q]} - \langle \log p \rangle_q$$

Of in the continuous case

$$KL[q||p] = \int q(x) \log \frac{q(x)}{p(x)} dx$$

- Fact: *KL* is convex and ≥ 0 (by Jensen ineq)
- Fact: *KL* is zero if and only if $p = q$.

CONDITIONAL ENTROPY AND MUTUAL INFORMATION

- Conditional entropy is defined as

$$H[p(x|y)] = -\int\int p(x,y)\log p(x|y)dxdy = H[p(x,y)] - H[p(y)]$$

  and captures the residual uncertainty on x once y is known.

- Mutual information between r.v. x and y is defined as

$$I[x,y] = KL[p(x,y)|p(x)p(y)] = H[p(x)] - H[p(x|y)]$$

  and captures the reduction in uncertainty about x by knowing y, i.e. it is a measure of how much y brings information about x, and viceversa.

Justification for maximum likelihood

$x_i \in \{0,1\}$

- Given a data set $\mathrm{x} = \{x_i\}, \quad i = 1, \ldots, N$, let the empirical distribution be

$$p_{emp}(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(x_i)$$

$$p_{emp}(0) = \frac{n_0}{N}$$

$$p_{emp}(1) = \frac{n_1}{N}$$

with $\mathbb{I}$ the indicator function of a set

- To find a suitable distribution $\boxed{q}$ to model the data, one may wish to minimize the Kullback-Leibler divergence

$$\underbrace{KL[\underset{\uparrow}{p_{emp}}||\underset{\uparrow}{q}]}_{} = H[\underset{\uparrow}{\cancel{p_{emp}}}] - \langle \log \underset{\uparrow}{q}(x)\rangle_{p_{emp}} = \overbrace{-\frac{1}{N}\sum \log q(x_i)}^{\text{log-likelih.}} + \text{const}$$

- *Maximum likelihood is equivalent to minimizing a KL divergence with the empirical distirbution*

OUTLINE

ORARIO

LUN	16-18	T
MAR	14-15.30	T
MER	14-17.00	L

1. BASICS OF PROBABILITY THEORY
2. SOME PROBABILITY DISTRIBUTIONS
3. FITTING DISTRIBUTIONS
4. INFORMATION THEORY
5. DECISION THEORY

(LUN 19/10 -> ORE 16.30)

AN OVERVIEW

- Suppose we have a classification problem, and we are able to learn a model of the joint distribution $p(x, y)$, where y is a categorical variable. Given a new input x^*, for which we want to make a prediction, to which class should we assign it?

AN OVERVIEW

- Suppose we have a classification problem, and we are able to learn a model of the joint distribution $p(x, y)$, where y is a categorical variable. Given a new input x^*, for which we want to make a prediction, to which class should we assign it?
- We may choose to assign it to class j if $p(y = j|x^*)$ is the maximum one. However, suppose y models having or not a cancer, and that $p(y = 0|x^*) = 0.51 > 0.49 = p(y = 1|x^*)$.

AN OVERVIEW

- Suppose we have a classification problem, and we are able to learn a model of the joint distribution $p(x, y)$, where y is a categorical variable. Given a new input x^*, for which we want to make a prediction, to which class should we assign it?
- We may choose to assign it to class j if $p(y = j|x^*)$ is the maximum one. However, suppose y models having or not a cancer, and that
$p(y = 0|x^*) = 0.51 > 0.49 = p(y = 1|x^*)$.
- To be more flexible, we can specify a loss function (or utility function), which is the cost $c_{k,j}$ of assigning x^* to class j when the true class is k.

AN OVERVIEW

- Suppose we have a classification problem, and we are able to learn a model of the joint distribution $p(x, y)$, where y is a categorical variable. Given a new input x^*, for which we want to make a prediction, to which class should we assign it?
- We may choose to assign it to class j if $p(y = j|x^*)$ is the maximum one. However, suppose y models having or not a cancer, and that $p(y = 0|x^*) = 0.51 > 0.49 = p(y = 1|x^*)$.
- To be more flexible, we can specify a loss function (or utility function), which is the cost $c_{k,j}$ of assigning x^* to class j when the true class is k.
- Then we can assign a point x^* to the class j minimising the expected loss w.r.t. the learned joint distribution (i.e. $\sum_k c_{k,j} p(y = k|x^*)$).