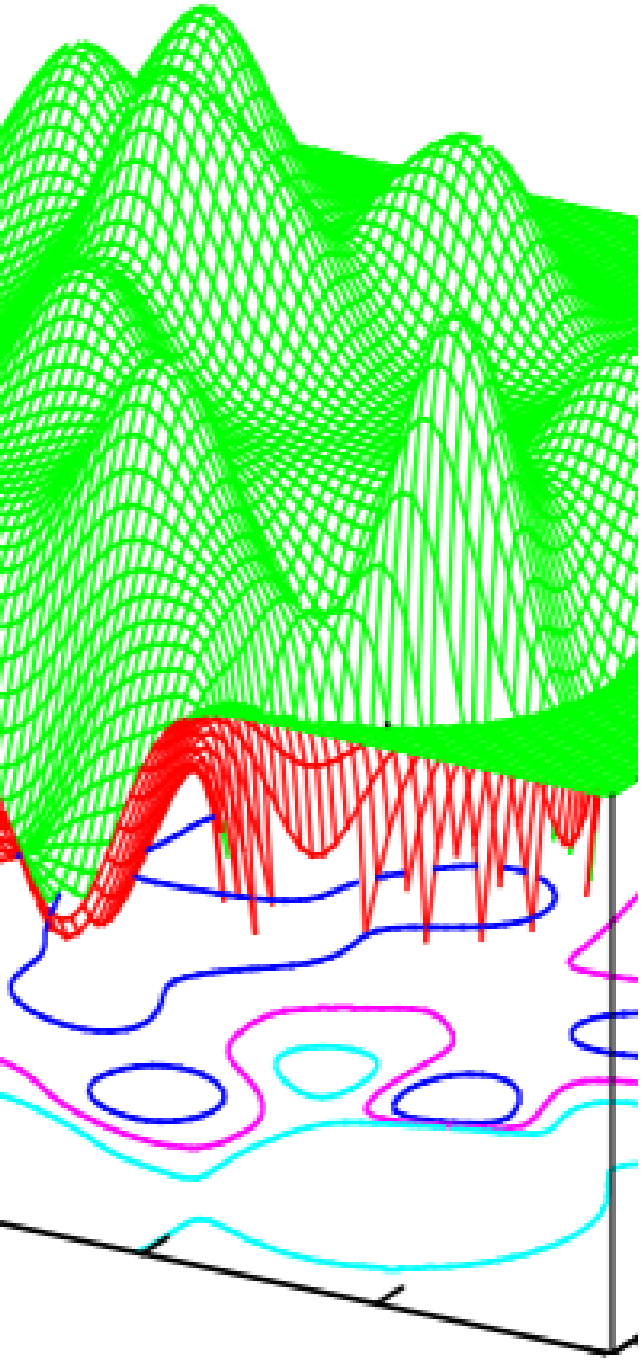


Machine Learning

Bayes Basics

Bayes, probabilities, Bayes' theorem & examples

Marc Toussaint
U Stuttgart

- So far:
  - Basic regression & classification methods:

    Features + Loss + Regularization & CV

  - All kinds of extensions & ideas to improve upon this

- Today: **Bayes**
  - A fully alternative framework to think about learning
  - A framework for modelling learning and inference problems in general
  - A framework to derive learning algorithms for specific models

The need for modelling

- Given a real world problem, translating it to a well-defined learning problem is non-trivial.

- The “framework” of plain regression/classification is rather restricted: input x, output y.

- Graphical models (probabilstic models with multiple random variables and dependencies) are a more general framework for modelling “problems”; regression & classification become a special case; Reinforcement Learning, decision making, unsupervised learning, but also language processing, image segmentation, are special cases.

Thomas Bayes (1702-1761)

Rev. T. Bayes

"Essay Towards Solving a Problem in the Doctrine of Chances"

Thomas Bayes (1702-1761)

Rev. T. Bayes

"Essay Towards Solving a Problem in the Doctrine of Chances"

- Addresses problem of *inverse probabilities*:
  Knowing the conditional probability of B given A, what is the conditional probability of A given B?

- Example:
  40% Bavarians speak dialect, only 1% of non-Bavarians speak (Bav.) dialect
  Given a random German that speaks non-dialect, is he Bavarian?
  (15% of Germans are Bavarian)

- “Inference” = Given some pieces of information (prior, observed variabes) what is the implication (the implied information, the posterior) on a non-observed variable

- "Inference" = Given some pieces of information (prior, observed variabes) what is the implication (the implied information, the posterior) on a non-observed variable

Learning as Inference

– given pieces of information: data, assumed model, *prior* over β
– non-observed variable: β

Probability Theory

- Why do we need probabilities?

Probability Theory

- Why do we need probabilities?

  – Obvious: to express inherent stochasticity of the world (data)

Probability Theory

- Why do we need probabilities?

  – Obvious: to express inherent stochasticity of the world (data)

- But beyond this: (also in a "deterministic world"):
  – lack of knowledge!
  – hidden (latent) variables
  – expressing *uncertainty*
  – expressing *information* (and lack of information)

- Probability Theory: an information calculus

Probability: Frequentist and Bayesian

- Frequentist probabilities are defined in the limit of an infinite number of trials
  Example: “The probability of a particular coin landing heads up is 0.43”

- Bayesian (subjective) probabilities quantify degrees of belief
  Example: “The probability of it raining tomorrow is 0.3”
  – Not possible to repeat “tomorrow”

Probabilities & Random Variables

- For a random variable X with discrete domain $\text{dom}(X) = \Omega$ we write:
  $\forall_{x \in \Omega} : \ 0 \leq P(X = x) \leq 1$
  $\sum_{x \in \Omega} P(X = x) = 1$

  Example: A dice can take values $\Omega = \{1, .., 6\}$.
  X is the random variable of a dice throw.
  $P(X = 1) \in [0, 1]$ is the probability that X takes value 1.

Probabilities & Random Variables

- For a random variable X with discrete domain $\text{dom}(X) = \Omega$ we write:
  $\forall_{x \in \Omega} : \ 0 \leq P(X = x) \leq 1$
  $\sum_{x \in \Omega} P(X = x) = 1$

  Example: A dice can take values $\Omega = \{1, .., 6\}$.
  X is the random variable of a dice throw.
  $P(X = 1) \in [0, 1]$ is the probability that X takes value 1.

- A bit more formally: a random variable relates a measureable space with a domain (sample space) and thereby introduces a probability measure on the domain ("assigns a probability to each possible value")

Probabilty Distributions

- $P(X=1) \in \mathbb{R}$ denotes a specific probability
  $P(X)$ denotes the probability distribution (function over Ω)

Probabilty Distributions

- $P(X=1) \in \mathbb{R}$ denotes a specific probability
  $P(X)$ denotes the probability distribution (function over Ω)

  Example: A dice can take values $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  By $P(X)$ we discribe the full distribution over possible values $\{1, .., 6\}$. These are 6 numbers that sum to one, usually stored in a *table*, e.g.: $[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$

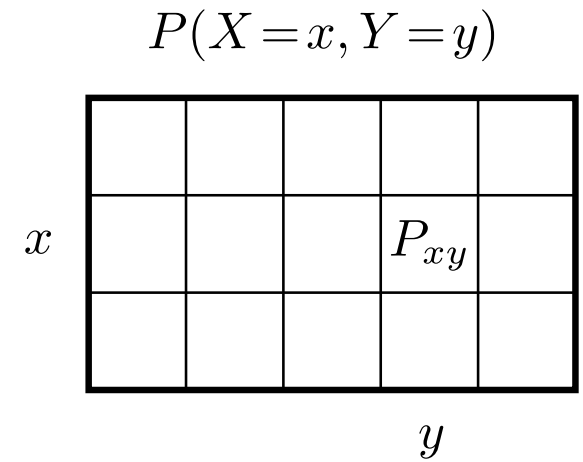
- In implementations we typically represent distributions over discrete random variables as tables (arrays) of numbers

- Notation for summing over a RV:
  In equation we often need to sum over RVs. We then write
  $$\sum_X P(X) \ \cdots$$
  as shorthand for the explicit notation $\sum_{x \in \text{dom}(X)} P(X=x) \ \cdots$

Joint distributions

Assume we have *two* random variables X and Y

- Definitions:
  Joint: $P(X, Y)$
  Marginal: $P(X) = \sum_Y P(X, Y)$
  Conditional: $P(X|Y) = \frac{P(X,Y)}{P(Y)}$



The conditional is normalized: $\forall_Y : \sum_X P(X|Y) = 1$

- X is *independent* of Y iff: $P(X|Y) = P(X)$
  (table thinking: all columns of $P(X|Y)$ are equal)

Joint distributions

joint: $P(X, Y)$
marginal: $P(X) = \sum_Y P(X, Y)$
conditional: $P(X|Y) = \frac{P(X,Y)}{P(Y)}$

- Implications of these definitions:
  product rule: $P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$

  Bayes' Theorem $P(X|Y) = \frac{P(Y|X)\ P(X)}{P(Y)}$

Bayes' Theorem

$$P(X|Y) = \frac{P(Y|X)\ P(X)}{P(Y)}$$

$$\text{posterior} = \frac{\text{likelihood} \cdot \text{prior}}{\text{normalization}}$$

Example 1: Bavarian dialect

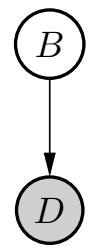
- 40% Bavarians speak dialect, only 1% of non-Bavarians speak (Bav.) dialect
  Given a random German that speaks non-dialect, is he Bavarian?
  (15% of Germans are Bavarian)

  $P(D=1 \mid B=1) = 0.4$
  $P(D=1 \mid B=0) = 0.01$
  $P(B=1) = 0.15$

Example 1: Bavarian dialect

- 40% Bavarians speak dialect, only 1% of non-Bavarians speak (Bav.) dialect
  Given a random German that speaks non-dialect, is he Bavarian?
  (15% of Germans are Bavarian)

$$P(D=1 \mid B=1) = 0.4$$
$$P(D=1 \mid B=0) = 0.01$$
$$P(B=1) = 0.15$$



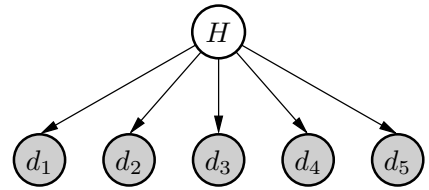
If follows

$$P(B=1 \mid D=0) = \frac{P(D=0 \mid B=1)\ P(B=1)}{P(D=0)} = \frac{.6 \cdot .15}{.6 \cdot .15 + 0.99 \cdot .85} \approx 0.097$$

Example 2: Coin flipping

HHTHT

HHHHH



- What process produces these sequences?

- We compare two hypothesis:
  $H = 1$: fair coin $\quad P(d_i = \mathtt{H} \mid H = 1) = \frac{1}{2}$
  $H = 2$: always heads coin $\quad P(d_i = \mathtt{H} \mid H = 2) = 1$

- Bayes' theorem:

$$P(H \mid D) = \frac{P(D \mid H) P(H)}{P(D)}$$

Coin flipping

$D = \texttt{HHTHT}$

$$P(D \mid H=1) = 1/2^5 \qquad P(H=1) = \tfrac{999}{1000}$$
$$P(D \mid H=2) = 0 \qquad P(H=2) = \tfrac{1}{1000}$$

$$\frac{P(H=1 \mid D)}{P(H=2 \mid D)} = \frac{P(D \mid H=1)}{P(D \mid H=2)} \frac{P(H=1)}{P(H=2)} = \frac{1/32}{0} \frac{999}{1} = \infty$$

Coin flipping

$D = \texttt{HHHHH}$

$$P(D \mid H=1) = 1/2^5 \qquad P(H=1) = \tfrac{999}{1000}$$
$$P(D \mid H=2) = 1 \qquad P(H=2) = \tfrac{1}{1000}$$

$$\frac{P(H=1 \mid D)}{P(H=2 \mid D)} = \frac{P(D \mid H=1)}{P(D \mid H=2)} \frac{P(H=1)}{P(H=2)} = \frac{1/32}{1} \frac{999}{1} \approx 30$$

Coin flipping

$D = \texttt{HHHHHHHHHH}$

$$P(D \mid H=1) = 1/2^{10} \qquad P(H=1) = \tfrac{999}{1000}$$
$$P(D \mid H=2) = 1 \qquad P(H=2) = \tfrac{1}{1000}$$

$$\frac{P(H=1 \mid D)}{P(H=2 \mid D)} = \frac{P(D \mid H=1)}{P(D \mid H=2)} \frac{P(H=1)}{P(H=2)} = \frac{1/1024}{1} \frac{999}{1} \approx 1$$

Learning as Bayesian inference

- *Think big:*

$$P(\text{World}|\text{Data}) = \frac{P(\text{Data}|\text{World})\ P(\text{World})}{P(\text{Data})}$$

$P(\text{World})$ describes our prior over all possible worlds. Learning means to infer about the world we live in based on the data we have!

Learning as Bayesian inference

- *Think big:*

$$P(\text{World}|\text{Data}) = \frac{P(\text{Data}|\text{World})\ P(\text{World})}{P(\text{Data})}$$

$P(\text{World})$ describes our prior over all possible worlds. Learning means to infer about the world we live in based on the data we have!

- In the context of regression, the "world" is the function $f(x)$

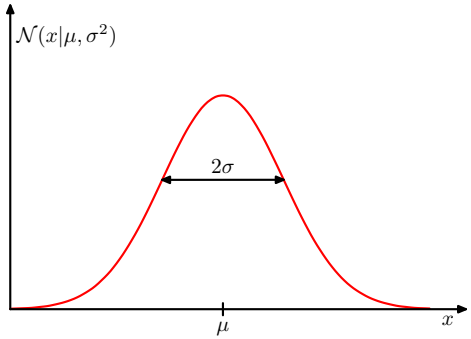
$$P(f|\text{Data}) = \frac{P(\text{Data}|f)\ P(f)}{P(\text{Data})}$$

$P(f)$ describes our prior over possible functions

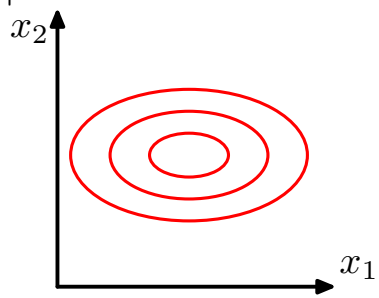
Regression means to infer the function based on the data we have

Gaussian distribution

- 1-dim: $\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{|2\pi\sigma^2|^{1/2}} \; e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$



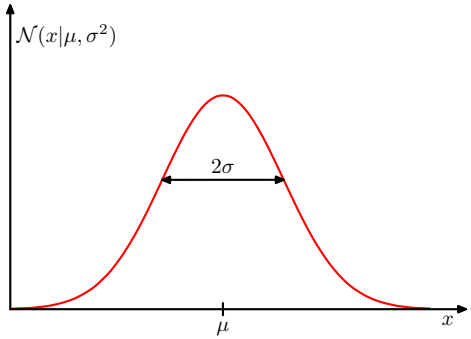
- n-dim: $\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \; e^{-\frac{1}{2}(x-\mu)^\top \; \Sigma^{-1} \; (x-\mu)} \propto e^{-\frac{1}{2}[x^\top\Sigma^{-1}x - 2x^\top\Sigma^{-1}\mu]}$



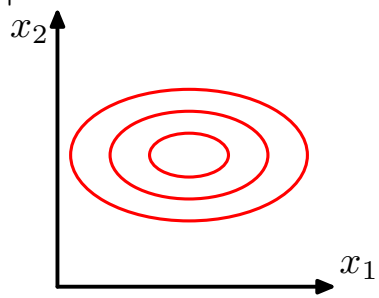
(b)

Gaussian distribution

- 1-dim: $\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{|2\pi\sigma^2|^{1/2}} \; e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}$



- n-dim: $\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \; e^{-\frac{1}{2}(x-\mu)^\top \; \Sigma^{-1} \; (x-\mu)} \propto e^{-\frac{1}{2}[x^\top\Sigma^{-1}x - 2x^\top\Sigma^{-1}\mu]}$



(b)

- continuous domain: *probability distribution* $F(x) = \int_{-\infty}^{x} dx \; p(x) \in [0, 1]$ is the integral of a *probability density function* $p(x) \in [0, \infty)$
  discrete domain: *probability distribution* and *probability mass function* $P(X) \in [0, 1]$ are used synonymously