Machine Learning: Fundamentals

A **model** is a formal description of a belief about the world.

Learning is the construction and/or revision of a model in response to **observations** of the world.

The mathematical/statistical foundation of machine learning:

- **Bayesian inference**: how to learn from observations *(today)*
- **maximum likelihood**: how to quantify the fit between a model and observations *(next week)*
- **optimization**: how to improve the fit of your model
- **regularization, model selection**: what is the best model?

Bayesian Inference

Rev. Thomas Bayes (1702-1761)

What is a Probability?

Technical answer:
any number P that obeys the **axioms of probability**:

- must lie between zero and one: $0 \leq P \leq 1$
- must add up: $P(A + B) = P(A) + P(B) - P(AB)$
  ($P(AB) = 0$ when A and B are mutually exclusive)
- must sum to one for mutually exclusive and collectively exhaustive alternatives

$P(A + B) = P(\text{“A or B”})$
$P(AB) = P(A \times B) = P(\text{“A and B”})$
$P(\bar{A}) = P(\text{“not A”}) = 1 - P(A)$

Interpretations of Probability

There are two interpretations of probability in statistics:

- **frequentist**: probability is the limit of observed frequency as number of observations goes to infinity. It is purely descriptive.
  Example: *“70% of November days in Zürich are rainy”*
- **Bayesian**: probability is a “degree of confidence” that one attaches to an uncertain event (Bernoulli, 1654-1705). It can be manipulated, for instance by applying **Bayes’ Rule**.
  Example: *“there is a 30% chance of rain tomorrow”*

Both respect the axioms of probability. For machine learning the Bayesian view is crucial, since it allows us to update measures of belief (models!) in response to observations – that is, to *learn*.

Example: Coin Tosses

For a given coin, what is the probability of coming up heads?

- **frequentist**: toss the coin many times
  P(heads) ≈ #heads / #tosses (can also quantify uncertainty)

- **Bayesian**: probability is a measure of belief. Use **prior** knowledge of coins to initially assume P(heads) = 0.5; revise this model of the coin in response to observations if necessary. (**Bayes' rule** will tell us exactly how to do this.)

Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

P(A|B) is read "probability of A given B", meaning:
the probability of A, given that B has already occurred.

Example: *in a given country, 90% of households own a TV, and 54% own both a TV and VCR. What is the probability that a household with TV also owns a VCR?*

P(VCR | TV) = P(VCR × TV) / P(TV) = 0.54 / 0.9 = 0.6

Conditional Probability

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

We can rearrange the above to read:

$$P(AB) = P(A|B)\, P(B)$$

Note that if A and B are **independent**, we have

$$P(AB) = P(A)\, P(B), \quad \text{so} \quad P(A|B) = P(A).$$

Furthermore we always have

$$P(AB) = P(BA)$$

$$\Rightarrow \quad P(A|B)\, P(B) = P(B|A)\, P(A)$$

Derivation of Bayes' Rule

$$P(A|B)\, P(B) = P(B|A)\, P(A)$$

$$\Rightarrow P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Since $P(B) = P(BA) + P(B\bar{A}) = P(B|A)\, P(A) + P(B|\bar{A})\, P(\bar{A})$,

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B \mid A)\, P(A) + P(B \mid \overline{A})\, P(\overline{A})}$$

This is the simplest form of **Bayes' Rule**.

Bayes in Action: Coin Tosses

Let A = "coin is fair", $\bar{A}$ = "coin is double-headed"
(let's ignore other possibilities for now).

Assume **prior:** P(A) = 0.9

- First coin toss: heads (H)
  P(H|A) P(A) = 0.5 · 0.9 = 0.45
  P(H|$\bar{A}$) P($\bar{A}$) = 1.0 · 0.1 = 0.1

$$P(A \mid H) = \frac{P(H \mid A)\,P(A)}{P(H \mid A)\,P(A) + P(H \mid \bar{A})\,P(\bar{A})} = \frac{0.45}{0.45 + 0.1} \approx 0.82$$

Note how we have used the observation H to update our model (that is, belief in the fairness) of the coin.

More Coin Tosses

- Second coin toss: heads again! (HH)
  P(HH|A) P(A) = 0.25 · 0.9 = 0.225
  P(HH|$\bar{A}$) P($\bar{A}$) = 1.0 · 0.1 = 0.1

$$P(A \mid HH) = \frac{0.225}{0.225 + 0.1} \approx 0.69$$

third toss:

P(A|HHH) ≈ 0.53

- after the 4th head in a row, we are more inclined to think the coin is double-headed than fair:

P(HHHH|A) P(A) = 0.0625 · 0.9 = 0.05625

$$P(A \mid HHHH) = \frac{0.05625}{0.05625 + 0.1} = 0.36$$

Faith Restored?

- a weaker prior - say, P(A) = 0.8 - would have caused us to lose faith in the coin sooner: P(A|HH) = 0.5. Conversely, a stronger prior would have caused us to hold out longer.
- Even after a million "heads" though, a single "tail" suffices to restore our faith in the fairness of the coin:

$$P(HHH\ldots T|\bar{A})\,P(\bar{A}) = 0 \cdot 0.1 = 0 \;\Rightarrow\; P(A|HHH\ldots T) = 1$$

If this seems inappropriate, it indicates that our **prior** did not capture our actual prior knowledge of coins: we should allow for double-tailed or bent coins, lying lecturers, *etc.*

Bayes' Rule: General Form

Posterior belief given observations

Likelihood of observations given the model

Prior belief

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

Evidence: used as a normalization factor

When used for inference:

A: "Annahmen" (model)

B: "Beobachtungen" (data)

Using the Posterior

The posterior can be used

- to make **predictions**: example – what is the probability of "heads" on the third coin toss, given that "heads" came up twice before already?

$$P(H|HH) = P(H|HH,A)\,P(A|HH) + P(H|HH,\bar{A})\,P(\bar{A}\,|HH)$$
$$= 0.5 \cdot 0.69 + 1.0 \cdot (1.0 - 0.69) = 0.655$$

- to make **decisions**: example – after "heads" comes up for the third time, is the coin fair? Answer: **yes**, since

$$P(A|HHH) = 0.53, \quad \text{but} \quad P(\bar{A}|HHH) = 1.0 - 0.53 = 0.47$$

(When forced to decide: minimize the risk of being wrong by picking the alternative with the highest posterior.)

The Normalization Factor

- $P(B)$ is the same for all alternatives $A_k \Rightarrow$ doesn't affect which is best $\Rightarrow$ can be ignored for decision problems

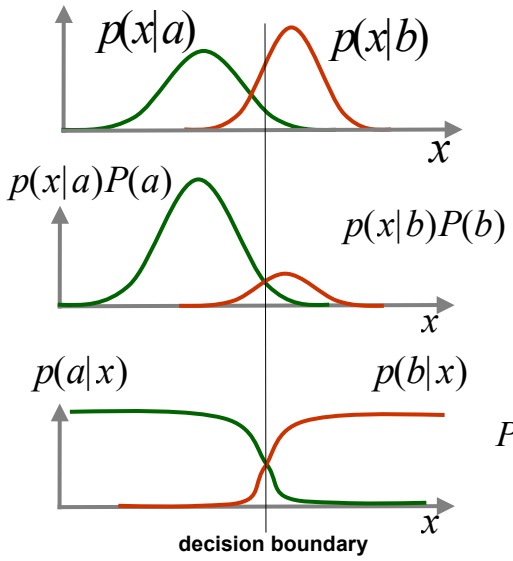
Otherwise:

- for *discrete*, mutually exclusive and exhaustive alternatives A_k: $P(B) = \sum_k P(B|A_k)\,P(A_k)$

- for a *continuous* spectrum of alternatives A:

$$P(B) = \int P(B|A)\,P(A)\,dA$$

Real-Valued Observations



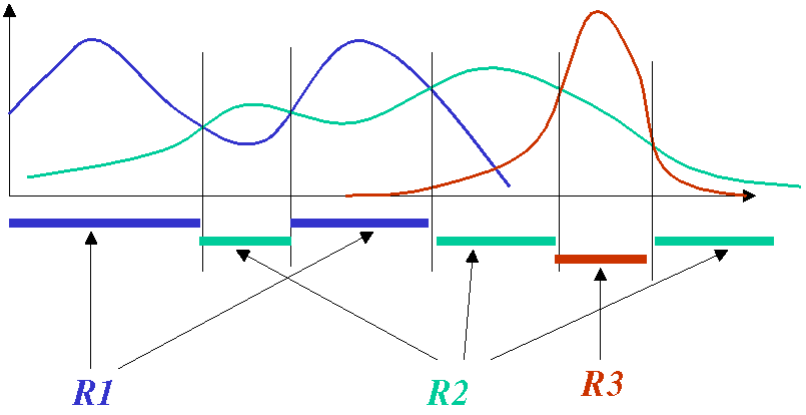
$$Likelihood$$

$$Likelihood \times Prior$$

$$Posterior = \frac{Likelihood \times Prior}{Normalization\ factor}$$

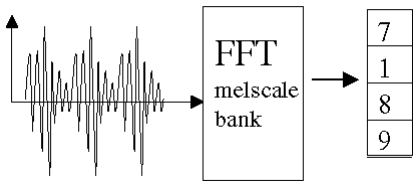
More Than Two Alternatives

- decision regions R_1, R_2, R_3, ...



Ambiguous Observations

- example: speech recognition system



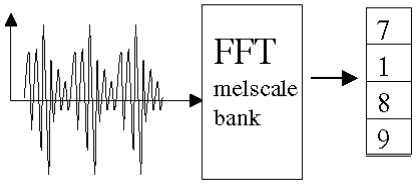
C_1: "This machine can recognize speech"

C_2: "This machine can wreck a nice beach"

- Both sentences sound the same. How should we decide?

Use Prior to Disambiguate

- Be smart about your priors! Here: use a language model.



$= x$

$p(x|C_k)$

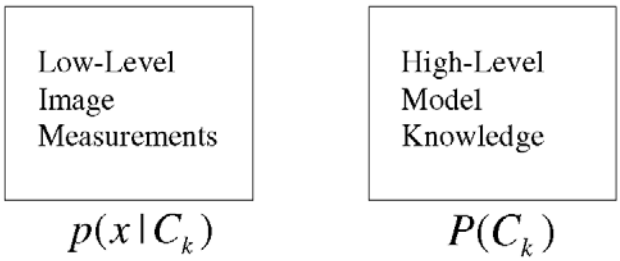
Language Model

P("wreck a nice beach") = 0.001
P("recognize speech") = 0.02

$P(C_k)$

Smart Priors

- another example: image recognition



- given your experiences in this lecture, how would a smart prior for coin tosses look like?

A Two-Stage Process

Bayesian inference thus comprises two equally important stages:

- **Construct the prior**: use all available prior knowledge to build a good model that lays out all plausible avenues for the inference machinery to explore.
- **Infer the posterior**: use Bayes' Rule to update the model in response to the available observations.

The same applies to **all** machine learning: finding a good way to incorporate prior knowledge is an important (though difficult, and often neglected) aspect of the problem.

Much of machine learning is in fact built upon the foundation of Bayesian inference – more next week.