Probability and Inference

Frequentist & Bayesian view

Oswin Krause, PML, Week 1 Tuesday

KØBENHAVNS UNIVERSITET

Notions of probability

What is the meaning of the word *probability* in the two statements below?

- Assume you throw a fair coin 100 times. What is the probability that the coin will come up heads more than 80 times?

- We throw a coin 100 times and it comes up heads 83 times. What is the probability that the coin is fair?

Notions of probability

What is the meaning of the word *probability* in the two statements below?

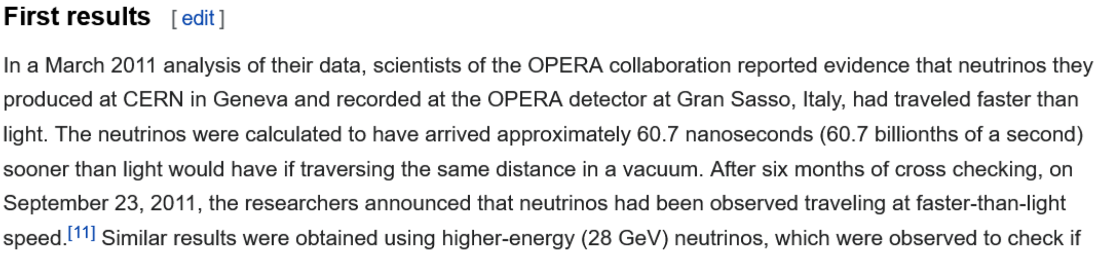
- Assume you throw a fair coin 100 times. What is the probability that the coin will come up heads more than 80 times?

→ Frequency of an outcome. *Statement about precision of an experiment.*

- We throw a coin 100 times and it comes up heads 83 times. What is the probability that the coin is fair?

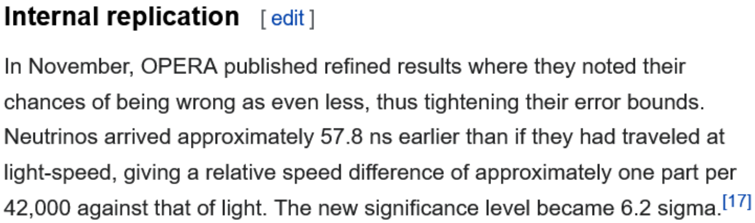
→ Reasoning about hypothesis. *Statement about likelihood of a belief.*

Example: Neutrinos travel faster than light?!?



First results [edit]

In a March 2011 analysis of their data, scientists of the OPERA collaboration reported evidence that neutrinos they produced at CERN in Geneva and recorded at the OPERA detector at Gran Sasso, Italy, had traveled faster than light. The neutrinos were calculated to have arrived approximately 60.7 nanoseconds (60.7 billionths of a second) sooner than light would have if traversing the same distance in a vacuum. After six months of cross checking, on September 23, 2011, the researchers announced that neutrinos had been observed traveling at faster-than-light speed.[11] Similar results were obtained using higher-energy (28 GeV) neutrinos, which were observed to check if



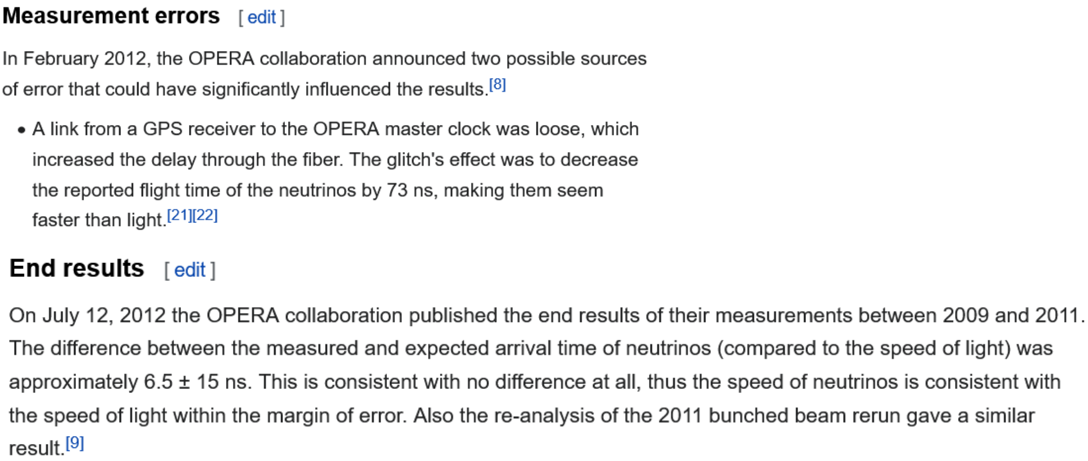
Internal replication [edit]

In November, OPERA published refined results where they noted their chances of being wrong as even less, thus tightening their error bounds. Neutrinos arrived approximately 57.8 ns earlier than if they had traveled at light-speed, giving a relative speed difference of approximately one part per 42,000 against that of light. The new significance level became 6.2 sigma.[17]

Example: Neutrinos travel faster than light?!?

- The experiment was very precise...
  - Signal to noise-ratio > 6 (6.2σ)
  - Probability of observing a random fluctuation this large was 1:10 million
- ...but did not change our beliefs
  - Most scientists believed in an error in the experiment
  - Overall, the experiment did not affect belief in faster-than-light Neutrinos

A loose cable...



Measurement errors [edit]

In February 2012, the OPERA collaboration announced two possible sources of error that could have significantly influenced the results.[8]

- A link from a GPS receiver to the OPERA master clock was loose, which increased the delay through the fiber. The glitch's effect was to decrease the reported flight time of the neutrinos by 73 ns, making them seem faster than light.[21][22]

End results [edit]

On July 12, 2012 the OPERA collaboration published the end results of their measurements between 2009 and 2011. The difference between the measured and expected arrival time of neutrinos (compared to the speed of light) was approximately 6.5 ± 15 ns. This is consistent with no difference at all, thus the speed of neutrinos is consistent with the speed of light within the margin of error. Also the re-analysis of the 2011 bunched beam rerun gave a similar result.[9]

Interpretations of probability theory

There are two common interpretations of probability

- Frequentist Probability
    - The probability of an event is its relative frequency if an experiment is repeated an infinite amount of times
    - Answers: How much does an estimate of an unknown parameter change when reproducing an experiment?
    - Assumption: Data is random, unknown parameter is fixed

Interpretations of probability theory

There are two common interpretations of probability

- Frequentist Probability
  - The probability of an event is its relative frequency if an experiment is repeated an infinite amount of times
  - Answers: How much does an estimate of an unknown parameter change when reproducing an experiment?
  - Assumption: Data is random, unknown parameter is fixed
- Bayesian Probability
  - Probability is the degree of belief about the state of the world
  - Answers: How much does my prior belief of the value of a parameter change after observing the outcome of an experiment?
  - Assumption: Data is fixed, unknown parameter is random.

Interpretations of probability theory

There are two common interpretations of probability

- Frequentist Probability
  - The probability of an event is its relative frequency if an experiment is repeated an infinite amount of times
  - Answers: How much does an estimate of an unknown parameter change when reproducing an experiment?
  - Assumption: Data is random, unknown parameter is fixed
- Bayesian Probability
  - Probability is the degree of belief about the state of the world
  - Answers: How much does my prior belief of the value of a parameter change after observing the outcome of an experiment?
  - Assumption: Data is fixed, unknown parameter is random.

Both approaches are consistent with probability theory.

Bayesian modeling

Bayesian modeling involves

- Unknown random parameter θ (e.g., model parameters)
- Dataset $\mathcal{D}$ (e.g.,input-label pairs in supervised learning)
- Prior distribution $p(\theta)$ (our belief which parameter is likely before seeing any data)
- Likelihood of $\mathcal{D}$ given θ, $p(\mathcal{D}|\theta)$

By Bayes' theorem, the posterior probability of θ given $\mathcal{D}$ is

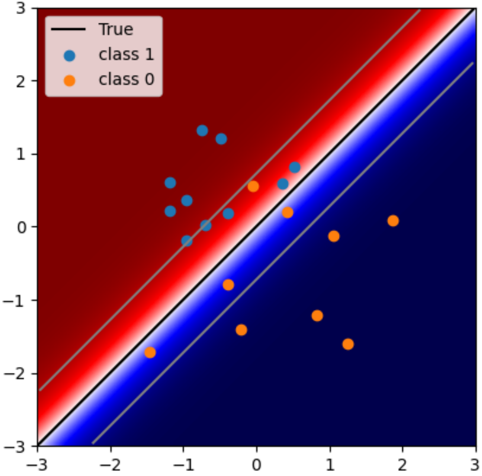
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta')p(\theta')d\theta'}$$

Example: Bayesian Logistic Regression

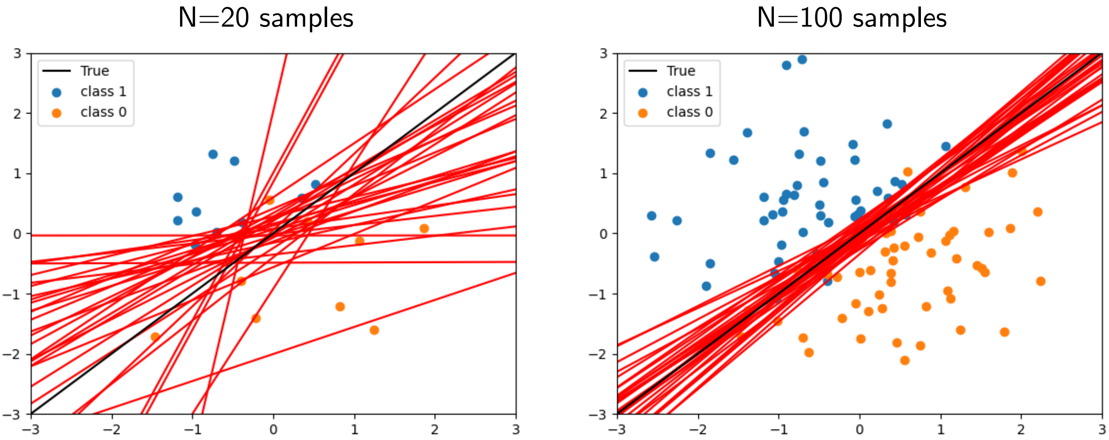
- $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$
  $y^{(i)} \in \{0, 1\}$
- $p(y = 1|x, \theta) = \text{sigmoid}(\theta_1 x_1 + \theta_2 x_2 + \theta_3)$
- $p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}, \theta)$
- Prior $p(\theta) = \frac{1}{\sqrt{2\pi}^3} \exp\left(-\frac{\|\theta\|^2}{2}\right)$

How does $p(\theta|\mathcal{D})$ look like?

Plot of True Model



Example Cont: Samples from Posterior $p(\theta|\mathcal{D})$

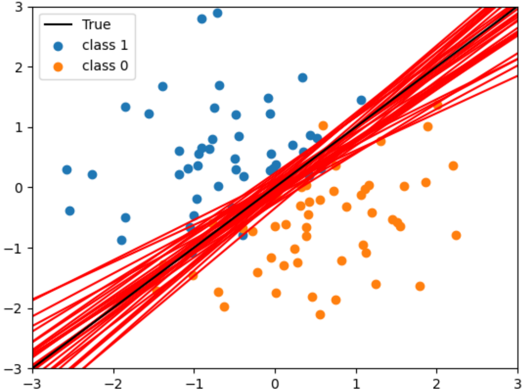


N=20 samples

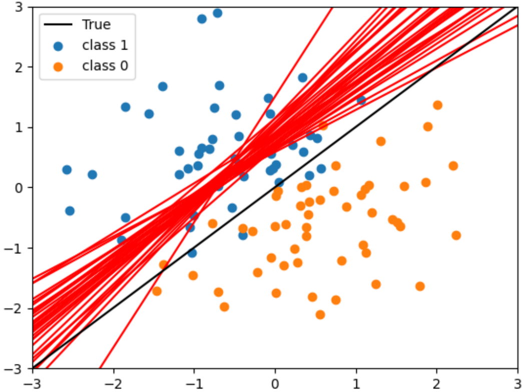
N=100 samples

Confidentially Wrong Prior = Confidentially wrong Posterior

$\theta_3 \sim \mathcal{N}(0,1)$



$\theta_3 \sim \mathcal{N}(5,0.1)$



Bayesian decision making

Bayesian methods give us a distribution over parameters, but how do we select the model? We can select the

- Mode (maximum a-posteriori, MAP)

$$\theta^* = \arg\max_{\theta} p(\theta|\mathcal{D})$$

- Mean parameter

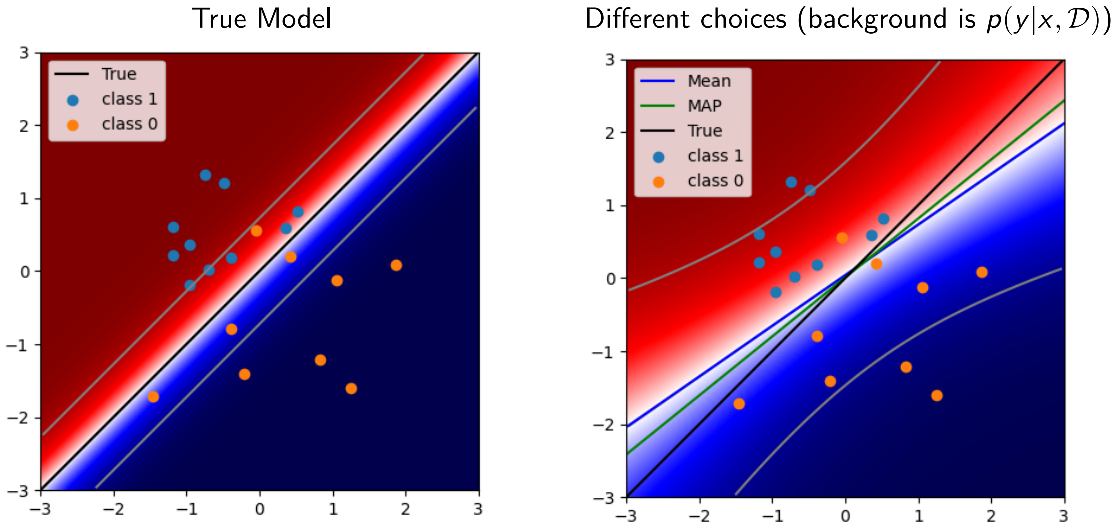
$$\hat{\theta} = \int p(\theta|\mathcal{D})\theta d\theta$$

- Posterior predictive

$$p(y|x,\mathcal{D}) = \int p(\theta|\mathcal{D})p(y|x,\theta)d\theta$$

Both mode or mean might not be representative of the distribution

Example: Bayesian Logistic Regression



True Model

Different choices (background is $p(y|x, \mathcal{D})$)

Bayesianism: Advantages and Shortcomings

Advantages

- Produces uncertainty estimates of model and predictions
- One unified approach (choose prior, model and likelihood → evaluate posterior)

Disadvantages

- No guarantees for model performance on new points
  - → Frequentist statistics on hold-out data
- Priors often unknown
  - → Weak priors (large variance on model parameters)
- Can not evaluate choice of prior
  - → Hierarchical priors (pick several prior candidates and assign prior probability to each of them.)
  - → Frequentist model selection (Empirical Bayes)

Frequentist modeling

Frequentist modeling involves

- Unknown fixed parameter θ^*
- Random data $\mathcal{D}$
- Estimator $\mathcal{A}$ producing estimate $\theta = \mathcal{A}(\mathcal{D})$ (can be deterministic)
- θ is a random variable depending on $\mathcal{A}$ and $\mathcal{D}$

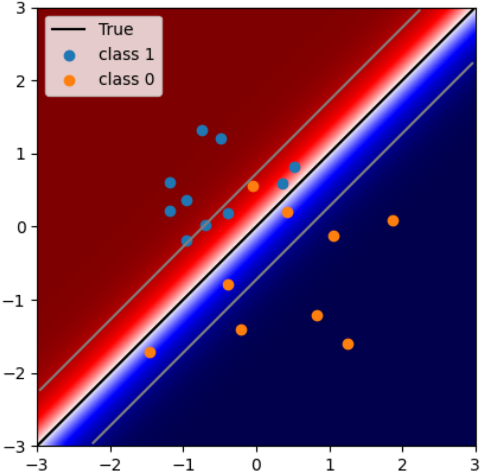
$$p(\theta|\theta^*, \mathcal{A}) = \int \underbrace{p(\mathcal{D}|\theta^*)}_{\substack{\text{Distribution of} \\ \text{datasets} \\ \text{with solution } \theta^*}} \quad \underbrace{p(\theta|\mathcal{A}, \mathcal{D})}_{\substack{\text{Distribution of} \\ \text{solutions by } A \\ \text{given } \mathcal{D}}} \; d\mathcal{D}$$

Example: (Frequentist) Logistic Regression

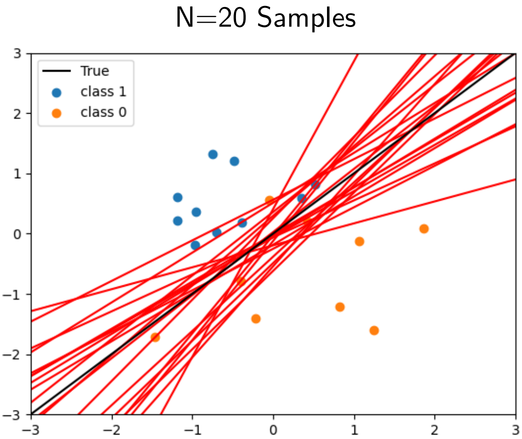
- $\mathcal{D} = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$
  $y^{(i)} \in \{0, 1\}$
- $p(y = 1|x, \theta) = \text{sigmoid}(\theta_1 x_1 + \theta_2 x_2 + \theta_3)$
- $p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}, \theta)$
- Estimator is maximum log-likelihood:

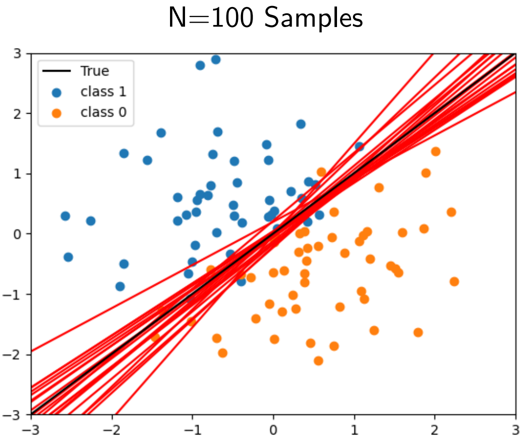
$$\mathcal{A}(\mathcal{D}) = \arg\max_{\theta} \log p(\mathcal{D}|\theta)$$

Plot of True Model



Example: Sample Distribution $p(\theta|\theta^*, \mathcal{A})$





Frequentist decision making

Importance of $p(\theta|\theta^*, \mathcal{A})$

- It is not a distribution over the likelihood of θ
- Contains only uncertainty of estimator and is no measure of quality.

→ Consider $\mathcal{A}(\mathcal{D}) = 0$

→ Or the MAP estimator using the bad prior shown earlier

We can't use $p(\theta|\theta^*, \mathcal{A})$ alone for evaluating estimators.

Frequentist decision making

What is a good estimator?

- Define the Risk of an estimator:

$$\mathcal{R}(\mathcal{A}, \theta^*) = \int p(\theta|\theta^*, \mathcal{A}) L(\theta, \theta^*) d\theta$$

  Here, $L(\theta, \theta^*)$ is a loss-function (e.g., squared loss)
- Measures expected loss for a single θ^* averaged over all $\mathcal{D}$
- A good estimator has low risk for all θ^*.

Frequentist decision making

We want a single number over all θ^*

- Minimax risk:
$$\mathcal{R}_{\mathsf{max}}(\mathcal{A}) = \max_{\theta^*} \mathcal{R}(\mathcal{A}, \theta^*)$$
This risk might be ∞
- Alternative: Weight risk using prior $p(\theta^*)$
- Bayes Risk
$$\mathcal{R}_{\mathsf{Bayes}}(\mathcal{A}) = \int p(\theta^*)\mathcal{R}(\mathcal{A}, \theta^*)d\theta^*$$
- Bayes risk is accepted among Frequentists if there is a *natural* choice of prior.

Example: Risk of a Classifier

Machine-Learning models are estimators

- Linear classifier $f(x) = \text{sign}(\phi^T x)$ (with fixed ϕ)
- Input: random data point x
- Output: estimate $y = f(x)$ of unknown fixed y^*

Example: Risk of a Classifier

The risk of a classifier can be defined as:

- Loss: $L(y, y^*) = \begin{cases} 0, & \text{if } y = y^* \\ 1, & \text{if } y \neq y^* \end{cases}$
- Risk: classification error of a specific class

$$\mathcal{R}(f, y^*) = \int p(x|y^*) L(f(x), y^*) \; dx$$

- Minimax Risk: Risk of the class with largest classification error
- Bayes risk with true sample distribution is classification loss

$$\mathcal{R}_{\text{Bayes}}(f) = \int p(x, y^*) L(f(x), y^*) \; dx \; dy^*$$

Frequentist model selection

Given two classifiers f_1, f_2 we want to pick the "better"

- Pick f_1 if $\mathcal{R}_{\text{Bayes}}(f_1) \leq \mathcal{R}_{\text{Bayes}}(f_2)$

→ But we often can't compute the integrals

- Estimate risk from data
- Empirical risk

$$\mathcal{R}_{\text{emp}}(f, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} L\left(f(x_T^{(i)}), y_T^{(i)}\right)$$

→ Use the model with lower empirical risk

- To claim "model performs better" we need to compute probability of difference in empirical risks

Frequentism: Advantages and Shortcomings

Advantages

- Can be applied to all Estimators
- Allows us to select models/priors

Disadvantages

- Risks are difficult to compute.
- Risk bounds are often pessimistic.
- Empirical risk measures introduce uncertainty into decision process
  (A lot of frequentist methodology is centered around developing statistical tests)