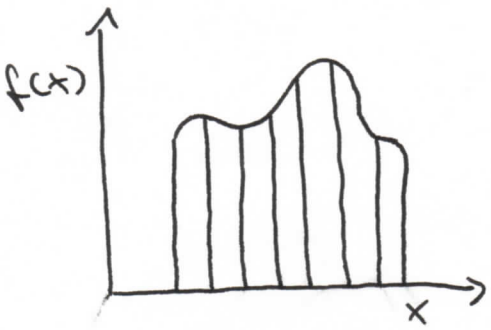
Lecture 05

Stratification:

$$I = \int_a^b f(x)\,dx = (b-a)\cdot \mathbb{E}[f(x)]$$

Idea: Split the domain into separate regions, take sample points from each region to estimate $\mathbb{E}[f]$ for this region and combine the results to get the final result.



M: number of strata: width $\frac{1}{M}$

N: total number of samples

L: number of samples per strata $L = \frac{N}{M}$

u_{ij}: i^{th} sample in the j^{th} strata, sampled from uniform distribution.

$$\mathbb{E}[f(x)] \approx \frac{1}{N} \sum_{ij} f(u_{ij}) \Rightarrow$$ forget the strata, just take all the points.

for each strata we have:

$$\mathbb{E}[f(u_j)] \approx \frac{1}{L} \sum_{i} f(u_{ij})$$

If we want to estimate $E[f(x)]$ using the intermediate results from all the strata we have:

$$E_S[f(x)] \approx \frac{1}{M} \sum_j E[f(u_j)] \Rightarrow$$ just average the estimates from all the strata.

Is this really the same? Is $E_S[f(x)] = E[f(x)]$?

$$E_S[f(x)] \approx \frac{1}{M} \sum_j E[f(u_j)] = \frac{1}{M} \sum_j \frac{1}{L} \cdot \sum_i f(u_{ij})$$

$$= \frac{1}{ML} \sum_{ij} f(u_{ij}) = \frac{1}{N} \sum_{ij} f(u_{ij}) \approx E[f(x)]$$

So far so good, but does stratification help us to get better estimates? Once again let's look at the variance:

$$\text{var}(E[f]) = \underbrace{\frac{\text{var}(f)}{N}}_{\text{MC error as discussed before.}} = \frac{E[f^2] - \overbrace{E^2[f]}^{\text{This is just what we are estimating squared.}}}{N}$$

$$E[f^2] = \frac{1}{N}\sum_{ij} f^2(u_{ij}) = \frac{1}{N}\sum_j L\left[\frac{1}{L}\sum_i f^2(u_{ij})\right]$$

$$= \frac{L}{N}\sum_j \underbrace{E[f^2(u_j)]}_{\text{per strata}} = \frac{1}{M}\sum_j E[f^2(u_j)]$$

The variance per strata is:

$$\text{var}(f(u_j)) = E[f^2(u_j)] - E^2[f(u_j)]$$

$$\Rightarrow E[f^2(u_j)] = \text{var}(f(u_j)) + E^2[f(u_j)]$$

Now we put it all together:

$$\text{var}(E[f(u_{ij})]) = \frac{E[f^2(u_{ij})] - E^2[f(u_{ij})]}{N}$$

$$= \frac{\frac{1}{M}\sum_j E[f^2(u_j)] - E^2[f(u_{ij})]}{N}$$

$$= \frac{\frac{1}{M}\sum_j \left(\text{var}(f(u_j)) + E^2[f(u_j)]\right) - E^2[f(u_{ij})]}{N}$$

$$= \frac{1}{NM}\sum_j \text{var}(f(u_j)) + \frac{1}{N}\left(\frac{\sum_j E^2[f(u_j)]}{M} - E^2[f(u_{ij})]\right)$$

Now the variance for the stratified estimate:

$$\text{var}\left(E_S[f(U_{ij})]\right) = \text{var}\left(\frac{1}{M}\sum_j E[f(U_j)]\right)$$

$$= \frac{1}{M^2}\,\text{var}\left(\sum_j E[f(U_j)]\right)$$

$$= \frac{1}{M^2}\sum_j \underbrace{\text{var}\left(E[f(U_j)]\right)}_{\text{MC error for the strata}}$$

There are no covariate terms here, because the strata are independent.

$$= \frac{1}{M^2}\sum_j \frac{\text{var}(f(U_j))}{L}$$

$$= \frac{1}{M^2 L}\sum_j \text{var}(f(U_j)) = \frac{1}{MN}\sum_j \text{var}(f(U_j))$$

Our stratified sampling is beneficial if:

$$\text{var}\left(E[f(U_{ij})]\right) - \text{var}\left(E_S[f(U_{ij})]\right) > 0$$

From all of the above we infer that we need

$$\frac{1}{N}\left(\underbrace{\frac{\sum_j E^2[f(U_j)]}{M} - E^2[f(U_j)]}_{\text{Cauchy-Schwarz inequality}}\right) > 0$$

$\Uparrow$ doesn't matter

If we want to gain even more from the stratified sampling, we can adjust the numbes of samples per strata depending on $\text{var}(f(U_j))$. To estimate the variance per strata though we need to run pilot studies, which can be computationally expensive.

You can also try and sample multiple rounds, using the estimates from previous rounds to judge the numbes of samples for the next.

Lecture 05 - Bayesian Formalism Part I

So far we have introduced some Monte Carlo methods for integration and sampling. To show applications where these methods really shine we will talk now about Bayesian and Frequentist approaches. This will motivate us to dig even deeper and learn to draw from more difficult distributions, which will lead us to Markov Chain Monte Carlo!

Assume we have some data D

And a model with parameters Θ

How likely is the data some instantiation of the model?

$\Rightarrow$ Likelihood: $p(D|\Theta)$

There are two ways of thinking about this:

1. There is a true model with some fixed Θ generating the data, and variance is caused by measurement errors. $\Rightarrow$ Frequentist

2. The parameters Θ are a random variable, the data comes from any of the models described by Θ, and we can model our beliefs about Θ as a prior $p(\Theta)$ $\Rightarrow$ Bayesian.

Frequentist statistics :

- Assume the "true" parameter is θ^*
- given the data we estimate $\hat{\theta}$
- repeat experiment m times → $\hat{\theta}_1, \ldots, \hat{\theta}_m$
- For $m \to \infty$ we have $\mathbb{E}[\hat{\theta}] \to \theta^*$
- need multiple repeats of the experiment
  ↳ Frequentist

What if we have only one set of data?
↳ sample from empical distribution
↳ Bootstrap

Bootstrap:

- We have n data points
- we sample m data sets a n points from D with replacement.

So how do we estimate $\hat{\theta}$?

Maximum Likelihood Estimation:

$D = \{d_1, d_2, \ldots, d_n\}$

If the noise for each data point is independent:

$$p(D|\theta) = \prod_i p(d_i|\theta)$$

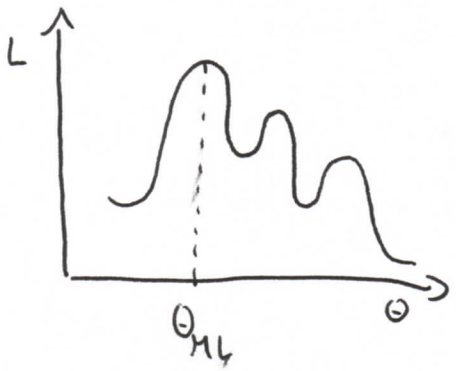
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta)$$

analytical solution:

- take derivative with respect to θ, set it to zero and find $\hat{\theta}_{ML}$
- Maximum of the likelihood is equal to the maximum of the log-likelihood
- Can also be numerically more stable
- Gets rid of the nasty product and makes it a nice sum

$$L(D|\theta) = -\log(p(D|\theta))$$
$$= -\sum_i \log(p(d_i|\theta))$$

numerical solution:

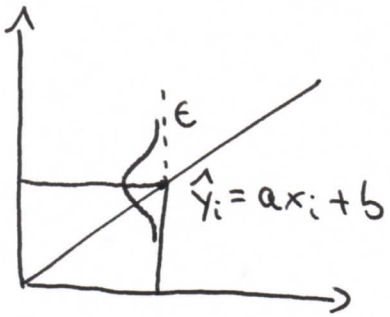


- sample, build histogram and find maximum
- Expensive in high dimensional spaces
- Can use tricks to find maximum without scanning the whole space.

Example: Linear regression

model: $y = ax + b + \epsilon$

- a: slope
- b: intercept
- ϵ: noise



$x = \{x_1, \dots, x_n\}$: Input, not random

$Y = \{y_1, \dots, y_n\}$: our noisy measurement data.

Since ϵ is Gaussian we expect a normal distribution around the true value of y_i.

$$p(d_i \mid \theta) \propto e^{-\frac{(y_i - (ax_i + b))^2}{2\sigma^2}}$$

↳ not normalized.

$$P(D \mid \theta) = \prod_i p(d_i \mid \theta)$$

$$L(D \mid \theta) = -\log(P(D \mid \theta)) \propto -\sum_i \frac{(y_i - (ax_i + b))^2}{2\sigma_\epsilon^2}$$

now all we need is

$$\frac{\partial L}{\partial \theta} \stackrel{!}{=} 0 \Rightarrow \frac{\partial L}{\partial a} \stackrel{!}{=} 0 \text{ and } \frac{\partial L}{\partial b} \stackrel{!}{=} 0$$

⇒ two equations, two unknowns

$$a_{ML} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad b_{ML} = \frac{1}{n}\left[\sum y_i - a_{ML}\sum_i x_i\right]$$

The difficulty is to find the right likelihood. Once you have it there are many ways to find θ^*.

Bayesian view:

- parameters θ live in probabilistic space.
- parameters do not have one value, they have a probability to have some value.
- We start with our belief what θ looks like
- Then we take the data and update our belief.
- The prior should not be motivated by the data, but by our knowledge about the world.

$$p(\theta|D) = \frac{p(D|\theta)\cdot p(\theta)}{p(D)}$$

- $p(\theta|D)$: posterior: new updated belief
- $p(D|\theta)$: likelihood
- $p(\theta)$: prior: our original belief.
- $p(D)$: evidence.

likelihood is what the data tells us about θ.

What if we don't know anything about the world?

- choose prior (θ) to be flat / uninformative

$\Rightarrow p(\theta|D) \propto p(D|\theta)$

↳ function of θ ↳ function of the data

Binomial example:

$p(D|\theta) \propto \theta^k (1-\theta)^{n-k}$ $\Rightarrow$ data is binomial

$p(\theta|D) \propto \theta^k (1-\theta)^{n-k}$ $\Rightarrow$ this is a function of θ, which is not binomial

Back to our posterior:

MAP: Maximum a posteriori estimate:

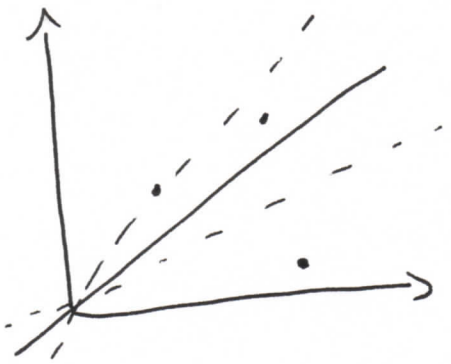
$$\theta_{MAP} = \underset{\theta}{\text{argmax}}\ P(\theta|D) = \underset{\theta}{\text{argmax}}\ p(D|\theta) \cdot p(\theta)$$

$\Rightarrow$ prior changes the result

$\Rightarrow$ prevents overfitting

$\Rightarrow$ Especially powerful for small data sets.

Let's look again at the regression example:



Outlier does not screw things up if we have a prior.

Model: $y = ax + b + \epsilon$

$\hookrightarrow N(0, \sigma_\epsilon)$

$\hookrightarrow b = 0$ to make things simpler

Likelihood:

$$p(D|\theta) \propto e^{\frac{-\sum_i (y_i - a x_i)^2}{2\sigma_\epsilon^2}}$$

Prior: $p(\theta): N(0, \sigma_\theta)$

posterior:

$$p(\theta|D) \propto e^{\frac{-\sum_i (y_i - a x_i)^2}{2\sigma_\epsilon^2}} \cdot e^{-\left(\frac{\theta^2}{2\sigma_\theta^2}\right)}$$

⇒ prior is Gaussian, and so is the likelihood

⇒ posterior is Gaussian too!

⇒ In this case we got lucky ⇒ or were clever planning ahead.

In real life multiplying the likelihood and prior can make things complicated

⇒ cannot find maximum analytically

⇒ cannot simply use scipy to draw from the posterior and build a histogram.

⇒ Monte Carlo to the rescue!