Data Science

Introduction to Machine Learning: Hypothesis Testing + Gradient Descent

April 12, 2021

Where were left off last time:

Preliminaries:

1. Different distributions
2. Different ways of reasoning about distributions (PDF, CDF)
3. Beginnings of Hypothesis Testing

Bounds

1. We discussed ways to use the CDF of a distribution to get *bounds* on some value
2. Without running more trials (or gathering more data), we can *increase* certainty by *widening* our bounds
3. But we weren't very concrete about how this relates to H_0 and H_1

Significance and Power

We need to talk about two aspects of interpreting experimental results:

1. *Significance*: How willing are we to reject H_0, even if it's true
2. *Power*: How willing are we to *fail* to reject H_0, even if it's false.

Errors

Significance and Power relate to errors.

1. Type 1 error: “false positive” (Significance)
2. Type 2 error: “false negative” (Power)

Errors in the Judicial System

	Innocent	Guilty
Guilty Verdict	???????	Correct
Not Guilty Verdict	Correct	??????

Errors in the Judicial System

	Innocent	Guilty
Guilty Verdict	Type 1	Correct
Not Guilty Verdict	Correct	??????

Errors in the Judicial System

	Innocent	Guilty
Guilty Verdict	Type 1	Correct
Not Guilty Verdict	Correct	Type 2

Back to our experiment (flipping a coin)

Our hypotheses:

1. H_0 the coin is fair ($p = 0.5$ that it lands Heads)
2. H_1 the coin is not fair ($p \neq 0.5$)

Back to our experiment (flipping a coin)

```
mu, sigma = normal_approx(1000, 0.5)
err = 0.05 # Our comfort with a type 1 error: 5%
lower, upper = norm_two_sided_bounds((1 - err), mu, sigma)
```

Back to our experiment (flipping a coin)

The result, with 95% probability:

1. Lower ≈ 469 result in heads
2. Upper ≈ 531 result in heads
3. What would we expect is the coin was fair?

Interpreting the results

Assuming the coin is fair

1. Just a 5% chance that the number of heads we'd see lies outside this range
2. Have we *proven* anything?
3. Are you convinced?
4. If you're wrong you lose a limb, are you convinced now?

Interpreting the results

But *we* got to choose the significance! How seriously should we take these results?

1. It is important that you communicate *why* you feel these results are valid.
2. It is *very easy* to lie with statistics:
    2.1 Imagine if H_0 was not in the 95% range, but in the 96% range
    2.2 Why is 5% special?

Communication, Communication, Communication

From an email I got last week (trying to book speakers):

```
I especially like your emphasis on communication
in data science.
```

p-Values

We computed *bounds* based on some chosen probability, *p-values* flips this around:

1. We assume H_0 is true.
2. We compute the probability that we would see a value *at least* as extreme as our actually observed value.

p-Values

Let's say we flipped a coin 1000 times (instead of having a distribution of such experiments)

1. We observe 530 heads, this would give us a p-value of 6.2%
2. We observe 532 heads, this would give us a p-value of 4.6%
3. (The function for computing the p-values is added to the notebook file)

Machine Learning

Many Machine Learning problems take the following form:

$$\text{minimize}_\theta \sum_{i=1}^{m} l(h_\theta(x^{(i)}), y^{(i)})$$

Let's go through this, bit by bit

Machine Learning

We have some input data we'd like to learn from:

$$\text{minimize}_\theta \sum_{i=1}^{m} l(h_\theta(x^{(i)}), y^{(i)})$$

Machine Learning

We have some known output data:

$$\text{minimize}_\theta \sum_{i=1}^{m} l(h_\theta(x^{(i)}), y^{(i)})$$

Machine Learning

We have a *hypothesis function*, with unkown parameters:

$$\text{minimize}_\theta \sum_{i=1}^{m} l(h_\theta(x^{(i)}), y^{(i)})$$

Machine Learning

We have a *loss* function that tells us how wrong we are:

$$\text{minimize}_\theta \sum_{i=1}^{m} l(h_\theta(x^{(i)}), y^{(i)})$$

Machine Learning

We want to sum the 'loss' from all of our input/output pairs:

$$\text{minimize}_\theta \sum_{i=1}^{m} l(h_\theta(x^{(i)}), y^{(i)})$$

Machine Learning

We want to minimize the 'loss' by changing the parameters to our hypothesis function:

$$\text{minimize}_\theta \sum_{i=1}^{m} l(h_\theta(x^{(i)}), y^{(i)})$$

One Approach

Gradient Descent!

1. The term gradient comes from calculus (a vector of partial derivatives)
2. We can ‘ride’ the gradient to some minimum (or maximum)

One Approach

Gradient Descent is search! The basic algorith:

1. pick a starting point
2. compute the sum of the loss over learning set
3. compute the sum of the loss for points ‘nearby’
4. pick new parameters based on the gradient from the previous steps
5. repeat
6. When do we stop?
7. What assumptions have we baked in?

Gradient Descent

Assumptions

1. That the loss function has a gadient!
2. That there's only one minimum (maximum)
3. What can we do about this?

Loss Functions

What we want:

1. Continuity
2. Global minimum
3. Cheap
4. Convex (why?)
5. A function is convex if a lin between two points always lies above the function.

Loss Functions

What we have:

1. Almost none of these things.
2. Most functions don't have these nice properties
3. Instead we *approximate* the loss function

Surrogate Loss Functions

Let's just make a function with the properties we care about!
Some alternatives:

1. 0/1 Loss
2. Hinge
3. Exponential
4. Squared Loss (very common)

On Wednesday we will:

1. Show examples of each loss function
2. Use Gradient descent to learn a linear model
3. Use our hypothesis testing to see if it's any good!

Thanks for your time!

:)