PRIORS

David Kauchak
CS159 Spring 2019

Admin

Assignment 7 due Friday at 5pm

Project proposals presentations at the beginning of class on Wednesday

Maximum likelihood estimation

Intuitive

Sets the probabilities so as to maximize the probability of the training data

Problems?

- Overfitting!
- Amount of data
  - particularly problematic for rare events
- Is our training data representative

Basic steps for probabilistic modeling

	Probabilistic models
Step 1: pick a model	Which model do we use, i.e. how do we calculate p(*feature, label*)?
Step 2: figure out how to estimate the probabilities for the model	How do train the model, i.e. how to we we estimate the probabilities for the model?
Step 3 (optional): deal with overfitting	How do we deal with overfitting?

Priors

Coin1 data: 3 Heads and 1 Tail

Coin2 data: 30 Heads and 10 tails

Coin3 data: 2 Tails

Coin4 data: 497 Heads and 503 tails

If someone asked you what the probability of heads was for each of these coins, what would you say?

Training revisited

From a probability standpoint, MLE training is selecting the Θ that maximizes:

$$p(\theta \mid data)$$

i.e.

$$\text{argmax}_\theta \, p(\theta \mid data)$$

We pick the most likely model parameters given the data

Estimating revisited

We can incorporate a prior belief in what the probabilities might be!

To do this, we need to break down our probability

$$p(\theta \mid data) = ?$$

(Hint: Bayes rule)

Estimating revisited

What are each of these probabilities?

$$p(\theta \mid data) = \frac{p(data \mid \theta) p(\theta)}{p(data)}$$

Priors

likelihood of the data under the model

probability of different parameters, call the prior

$$p(\theta \mid data) = \frac{p(data \mid \theta)p(\theta)}{p(data)}$$

probability of seeing the data (regardless of model)

Priors

$$\theta = \text{argmax}_\theta \frac{p(data \mid \theta)p(\theta)}{p(data)}$$

Does p(data) matter for the argmax?

Priors

likelihood of the data under the model

probability of different parameters, call the prior

$$\theta = \text{argmax}_\theta \, p(data \mid \theta)p(\theta)$$

What does MLE assume for a prior on the model parameters?

Priors

likelihood of the data under the model

probability of different parameters, call the prior

$$\theta = \text{argmax}_\theta \, p(data \mid \theta)p(\theta)$$

- Assumes a uniform prior, i.e. all Θ are equally likely!
- Relies solely on the likelihood

A better approach

$$\theta = \text{argmax}_\theta \; p(data \mid \theta) p(\theta)$$

$$likelihood(data) = \prod_{i=1}^{n} p_\theta(x_i)$$

We can use any distribution we'd like
This allows us to impart addition bias into the model

Another view on the prior

Remember, the max is the same if we take the log:

$$\theta = \text{argmax}_\theta \log(p(data \mid \theta)) + \log(p(\theta))$$

$$\log - likelihood = \sum_{i=1}^{n} \log(p(x_i))$$

We can use any distribution we'd like
This allows us to impart addition bias into the model

What about smoothing?

training data

for each label, pretend like we've seen each feature/word occur in λ additional examples

$$\theta_j = \frac{count(w_j, y)}{\sum_{k=1}^{m} count(w_k, y)}$$

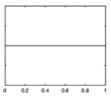
$$\theta_j = \frac{count(w_j, y) + \lambda}{\lambda m + \sum_{k=1}^{m} count(w_k, y)}$$

Sometimes this is also called smoothing because it is seen as smoothing or interpolating between the MLE and some other distribution

Prior for NB

$$\theta = \text{argmax}_\theta \log(p(data \mid \theta)) + \log(p(\theta))$$

Uniform prior



Dirichlet prior

$\lambda = 0$ → increasing

$$p(w_j \mid y) = \theta_j = \frac{count(w_j, y)}{\sum_{k=1}^{m} count(w_k, y)}$$

$$\theta_j = \frac{count(w_j, y) + \lambda}{\sum_{k=1}^{m} (count(w_k, y) + \lambda)} = \frac{count(w_j, y) + \lambda}{\lambda m + \sum_{k=1}^{m} count(w_k, y)}$$