Welcome to CS109A

Gili Rusak

Agenda

- Parameter Estimation*
- Beta Distribution*
- Naive Bayes Classifier*
- Applications

* Relevant for HW6

Parameter Estimation

Parameters and MLE

Suppose $x_1, \ldots, x_n$ are i.i.d. (independent and identically distributed) values sampled from some distribution with density function $f(x|\theta)$, where θ is unknown. Recall that the likelihood of the data is

$$L(\theta) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Recall we solve an optimization problem to find $\hat{\theta}$ which maximizes $L(\theta)$, i.e., $\hat{\theta} = \arg\max_\theta L(\theta)$.

1. Write an expression for the log-likelihood, $LL(\theta) = \log L(\theta)$.

2. Why can we optimize $LL(\theta)$ rather than $L(\theta)$?

3. Why do we optimize $LL(\theta)$ rather than $L(\theta)$?

Example Problem Solution

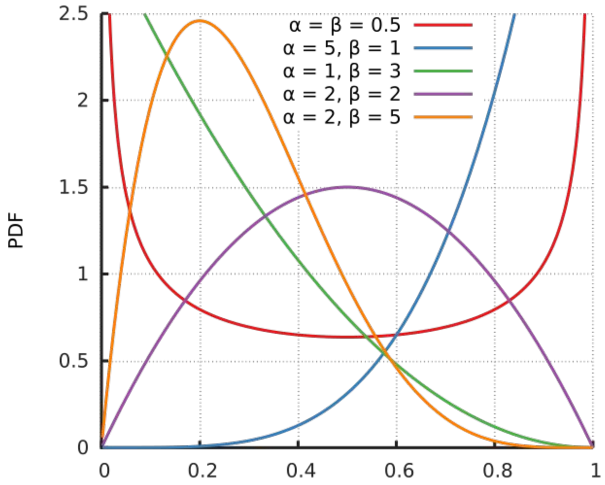
1. $LL(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$
2. The logarithm (for bases > 1) is a monotonically increasing function. This means that if $f(a) > f(b)$, then $\log(f(a)) > \log(f(b))$, so the arg max function is preserved across a logarithmic transformation, i.e., $\arg\max L(\theta) = \arg\max LL(\theta)$.
3. Logs turn products into sums, which makes taking the derivative for maximization or minimization much simpler.

Beta Random Variable

Beta Distribution

The Probability Density Function (PDF) for a Beta $X \sim Beta(a, b)$ is:

$$f(X = x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{where } B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$



Beta priors and posteriors for binomial random variables

1. Suppose you have a coin where you have no prior belief on its true probability of heads p. How can you model this belief as a Beta distribution?

2. Suppose you have a coin which you believe is fair, with "strength" α. That is, pretend you've seen α heads and α tails. How can you model this belief as a Beta distribution?

3. Now suppose you take the coin from the previous part and flip it 10 times. You see 8 heads and 2 tails. How can you model your posterior belief of the coin's probability of heads?

Beta Distribution Solution

1. Beta(1, 1) is a uniform prior, meaning that prior to seeing the experiment, all probabilities of heads are equally likely.
2. $\text{Beta}(\alpha + 1, \alpha + 1)$. This is our prior belief about the distribution.
3. $\text{Beta}(\alpha + 9, \alpha + 3)$

Naive Bayes Classifier

Classification Task

- Given a set of data about historical features, predict the label of a new set of features.
- Examples: given a set of cat and dog images. Build a model to predict whether a new image is a cat or a dog.

Naive Bayes Binary Classification Training

The objective in training is to estimate the probabilities $P(Y)$ and $P(X_j|Y)$ for all $0 < j \leq m$ features.
Using an MLE estimate:

$$\hat{P}(X_j = x_j|Y = y) = \frac{(\text{\# training examples where } X_j = x_j \text{ and } Y = y)}{(\text{training examples where } Y = y)}$$

Naive Bayes Binary Classification Training

The objective in training is to estimate the probabilities $P(Y)$ and $P(X_j|Y)$ for all $0 < j \leq m$ features.

Using an MLE estimate:

$$\hat{P}(X_j = x_j | Y = y) = \frac{(\text{\# training examples where } X_j = x_j \text{ and } Y = y)}{(\text{training examples where } Y = y)}$$

Using a Laplace MAP estimate:

$$\hat{P}(X_j = x_j | Y = y) = \frac{(\text{\# training examples where } X_j = x_j \text{ and } Y = y) + 1}{(\text{training examples where } Y = y) + 2}$$

Naive Bayes Binary Classification Training

The objective in training is to estimate the probabilities $P(Y)$ and $P(X_j|Y)$ for all $0 < j \leq m$ features.

Using an MLE estimate:

$$\hat{P}(X_j = x_j|Y = y) = \frac{(\text{\# training examples where } X_j = x_j \text{ and } Y = y)}{(\text{training examples where } Y = y)}$$

Using a Laplace MAP estimate:

$$\hat{P}(X_j = x_j|Y = y) = \frac{(\text{\# training examples where } X_j = x_j \text{ and } Y = y) + 1}{(\text{training examples where } Y = y) + 2}$$

Estimating $P(Y = y)$ is also straightforward. Using MLE estimation:

$$\hat{P}(Y = y) = \frac{(\text{\# training examples where } Y = y)}{(\text{training examples})}$$

Naive Bayes Binary Classification Prediction

For an example with $\mathbf{X} = [x_1, x_2, \ldots, x_m]$, we can make a corresponding prediction for Y. We use hats (e.g., $\hat{P}$ or $\hat{Y}$) to symbolize values which are estimated.

$$\hat{Y} = g(\mathbf{x}) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \, \hat{P}(Y)\hat{P}(\mathbf{X}|Y)$$

This is equal to argmax $\hat{P}(Y = y|\mathbf{X})$

$$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \, \hat{P}(Y = y) \prod_{j=1}^{m} \hat{P}(X_j = x_j | Y = y)$$

Naïve Bayes assumption

$$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \, \log \hat{P}(Y = y) + \sum_{j=1}^{m} \log \hat{P}(X_i = x_j | Y = y)$$

Log version for numerical stability

Naive Bayes Example

Say we have thirty examples of people's preferences (like or not) for Star Wars, Harry Potter and Pokemon. Each training example has X_1, X_2 and Y where X_1 is whether or not the user liked Star Wars, X_2 is whether or not the user liked Harry Potter and Y is whether or not the user liked Pokemon. For the 30 training examples, the MAP and MLE estimates are as follows:

Y \ X_1	0	1	MLE estimates	
0	3	10	0.23	0.77
1	4	13	0.24	0.76

Y \ X_2	0	1	MLE estimates	
0	5	8	0.38	0.62
1	7	10	0.41	0.59

Y	#	MLE est.
0	13	0.43
1	17	0.57

Naive Bayes Example

Say we have thirty examples of people's preferences (like or not) for Star Wars, Harry Potter and Pokemon. Each training example has X_1, X_2 and Y where X_1 is whether or not the user liked Star Wars, X_2 is whether or not the user liked Harry Potter and Y is whether or not the user liked Pokemon. For the 30 training examples, the MAP and MLE estimates are as follows:

Y \ X1	0	1	MLE estimates	
0	3	10	0.23	0.77
1	4	13	0.24	0.76

Y \ X2	0	1	MLE estimates	
0	5	8	0.38	0.62
1	7	10	0.41	0.59

Y	#	MLE est.
0	13	0.43
1	17	0.57

Y \ X1	0	1	MAP estimates	
0	3	10	0.27	0.73
1	4	13	0.26	0.74

Y \ X2	0	1	MAP estimates	
0	5	8	0.4	0.6
1	7	10	0.42	0.58

Y	#	MAP est.
0	13	0.45
1	17	0.55

Naive Bayes Example

For a new user who likes Star Wars ($X1 = 1$) but not Harry Potter ($X2 = 0$), do you predict that they will like Pokemon?

Naive Bayes Example

For a new user who likes Star Wars ($X_1 = 1$) but not Harry Potter ($X_2 = 0$), do you predict that they will like Pokemon? Yes! $Y = 1$ leads to a larger value in the argmax term:

$$\text{if } Y = 0 : \hat{P}(X_1 = 1|Y = 0)\hat{P}(X_2 = 0|Y = 0)\hat{P}(Y = 0) = (0.77)(0.38)(0.43) \approx 0.126$$

$$\text{if } Y = 1 : \hat{P}(X_1 = 1|Y = 1)\hat{P}(X_2 = 0|Y = 1)\hat{P}(Y = 1) = (0.76)(0.41)(0.57) \approx 0.178$$

Applications

- Classification problems occur in many disciplines
  - Computer Vision (CS131, CS231N)
  - Deep Learning (CS230)
  - Natural Language Processing (CS124, CS224N)
  - General Game Playing (CS227B)
  - Biocomputing (CS274)

See you next Tuesday!