10-701 INTRODUCTION TO MACHINE LEARNING (SCS MAJORS)
LECTURE 3: ESTIMATING PROBABILITIES

LEILA WEHBE
CARNEGIE MELLON UNIVERSITY
MACHINE LEARNING DEPARTMENT

LINKS (USE THE VERSION YOU NEED)

- Notebook (https://github.com/lwehbe/10701/blob/F22/Lecture_03_estimating_probabilities.ipynb)
- PDF slides (https://github.com/lwehbe/10701/raw/F22/Lecture_03_estimating_probabilities.pdf)

FUNCTION APPROXIMATION

Problem Setting:

- Set of possible instances X
- Unknown target function $f : X \to Y$
- Set of function hypotheses $H = \{h | h : X \to Y\}$

Input:

- Training examples $\{(x^{(i)}, y^{(i)})\}$ of unknown target function f

Output:

- Hypothesis $h \in H$ that best approximates target function f

FUNCTION APPROXIMATION - MORE FORMALLY

Problem Setting: Our training data is denoted:

$$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\} \subset \mathcal{R}^d \times \mathcal{C}$$

- $(\mathbf{x}_i, y_i)$ are sampled from an (unknown) distribution $P(X, Y)$.
- $\mathcal{R}$ is the feature space:
  - example attribute can be binary ($\mathcal{R} = \mathcal{N}^d$), real ($\mathcal{R} = \mathcal{R}^d$) or other
- $\mathcal{C}$ is the label space:
  - Binary Classification: $\mathcal{C}$ is $\{0, 1\}$ or $\{-1, 1\}$
  - Multiclass Classification: $\mathcal{C} = \{1, 2, \dots, K\}$
  - Regression: $\mathcal{C} = \mathbb{R}$

CHOOSING H

- Choosing the hypothesis class:
    - encodes important assumptions about the type of problem we are trying to learn
- No free lunch theorem (https://en.wikipedia.org/wiki/No_free_lunch_theorem):
    - every successful ML algorithm must make assumptions
    - no single ML algorithm that works for every setting

CHOOSING $h \in H$

- find h by choosing an appropriate loss function $\mathcal{L}(h)$ and finding the h^* that minimizes it.
- Examples of loss functions (we will see more in the course):
- Zero-one loss:
    - count how many mistakes
    - rarely used for optimization because not continuous - used to evaluate classifiers

$$\mathcal{L}_{0/1}(h) = \frac{1}{n}\sum_{i=1}^{n} I(h(\mathbf{x}_i) \neq y_i),$$

$$\text{where} \quad I(h(\mathbf{x}_i) \neq y_i) = \begin{cases} 1, & \text{if } h(\mathbf{x}_i) \neq y_i \\ 0, & \text{otherwise} \end{cases}$$

CHOOSING $h \in H$

- Squared loss:
    - typically used in regression settings
    - loss grows quadratically, encourages no predictions to be really far off
    - if prediction is really close, not encouraged to be exact

$$\mathcal{L}_{sq}(h) = \frac{1}{n}\sum_{i=1}^{n}(h(\mathbf{x}_i) - y_i)^2$$

CHOOSING $h \in H$

- Absolute loss:
    - typically used in regression settings
    - less sensitive to outliers / noisy points (grows linearly)

    - encourages exact solution

$$\mathcal{L}_{abs}(h) = \frac{1}{n} \sum_{i=1}^{n} |h(\mathbf{x}_i) - y_i|$$

TRAIN - TEST SPLITS

- Divide your data into training and test
    - never touch the test data!
- Validation
    - need to choose some hyperparameters. Don't use the test data!
    - Split data into Training - Validation - Test
    - Use validation to pick hyperparameters / know when to stop etc
- Cross-validation
    - divide the training set into parts and iteratively use one as validation, averaging the results across folds
    - LOOCV (leave one out cross validation), 10-fold cross-validation

TRAINING, TEST AND GENERALIZATION ERRORS

- learning:

$$h^* = \operatorname{argmin}_{h \in H} \frac{1}{|D_{train}|} \sum_{(\mathbf{x}, y) \in D_{train}} \ell(\mathbf{x}, y | h)$$

- evaluation:

$$\text{test error: } \epsilon_{\text{test}} = \frac{1}{|D_{test}|} \sum_{(\mathbf{x}, y) \in D_{test}} \ell(\mathbf{x}, y | h^*)$$

- if the data was drawn IID from P, then the testing loss is an unbiased estimate of the true generalization loss:

$$\text{generalization error: } \quad \epsilon = \mathbb{E}_{(\mathbf{x}, y) \sim P} [\ell(\mathbf{x}, y | h^*)]$$

BAYES OPTIMAL CLASSIFIER

- Example: Assume (and this is almost never the case) you knew $P(y|x)$, then you would simply predict the most likely label.
    - The Bayes optimal classifier predicts:

$$y^* = \operatorname{argmax}_y P(y|\mathbf{x})$$

- Although the Bayes optimal classifier is as good as it gets, it still can make mistakes. It is always wrong if a sample does not have the most likely label. We can compute the probability of that happening precisely (which is exactly the error rate):

$$\epsilon_{\text{BayesOpt}} = 1 - P(y^*|\mathbf{x})$$

BAYES OPTIMAL CLASSIFIER

- Why is the Bayes optimal classifier interesting, if it cannot be used in practice? The reason is that it provides a highly informative lower bound of the error rate. With the same feature representation no classifier can obtain a lower error.
- What to do when we don't have $P(y|x)$.

PROBABILISTIC FUNCTION APPROXIMATION:

Instead of $f : X \rightarrow Y$,
learn $P(Y|X)$

CONDITIONAL PROBABILITY

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

CONDITIONAL PROBABILITY

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

CORROLLARY: THE CHAIN RULE:

$$P(A, B) = P(A|B)P(B)$$

(VERY) QUICK RECAP

RANDOM VARIABLES AND PROBABILITY TABLES

Example:

- Experiment: flipping a coin
- Universe: {0 = Tails, 1 = Heads)
- Events: 0 (Tails) and 1 (Heads).
- A: random variable indicating that the coin is heads.

A	$P(A)$
0	0.8
1	0.2

JOINT DISTRIBUTION

We can have multiple random variables in the same experiment.

The universe corresponds to the set of all joint settings of the variables.

For example, if we throw the same coin three times, we can define:

- A: random variable that indicates that the first flip is heads.
- B: random variable that indicates that the second flip is heads.
- C: random variable that indicates that the third flip is heads.

One example of an event is: $A = 0 \wedge B = 1 \wedge C = 0$.

We can write a probability table listing all the configuration of values for the variables:

A	B	C	P
0	0	0	0.512
0	0	1	0.128
0	1	0	0.128
0	1	1	0.032
1	0	0	0.128
1	0	1	0.032
1	1	0	0.032
1	1	1	0.008

We can write a probability table listing all the configuration of values for the variables:

A	B	C	P
0	0	0	0.512
0	0	1	0.128
0	1	0	0.128
0	1	1	0.032
1	0	0	0.128
1	0	1	0.032
1	1	0	0.032
1	1	1	0.008

In this coin example, the variables are independent from each other

(you can see for yourself that $P(A|B,C) = P(A)$).

We can consider another example: we select a day at random:

- A: the sky is cloudy.
- B: the temperature is cold.
- C: it rains.

A	B	C	P
0	0	0	0.3
0	0	1	0.1
0	1	0	0.15
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.1
1	1	1	0.15

JOINT DISTRIBUTION

A − cloudy	B − cold	C − rain	P
0	0	0	0.3
0	0	1	0.1
0	1	0	0.15
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.1
1	1	1	0.15

We can use the joint distribution to compute the probability of any compound event:

$$P(E) = \sum_{\text{row matching E}} P(\text{row})$$

What is the probability of the day being not cloudy and cold?

JOINT DISTRIBUTION

A − cloudy	B − cold	C − rain	P
0	0	0	0.3
0	0	1	0.1
0	1	0	0.15
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.1
1	1	1	0.15

What is the probability of the day being not cloudy and cold?

$$P(A = 0, B = 1) = P(A = 0, B = 1, C = 0) + P(A = 0, B = 1, C = 1) = 0.2$$

We can use the joint probability table to marginalize variables out

$A - \text{cloudy}$	$B - \text{cold}$	$C - \text{rain}$	P
0	0	0	0.3
0	0	1	0.1
0	1	0	0.15
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.1
1	1	1	0.15

What is:

A	$P(A)$
0	
1	

?

$A -$ cloudy	$B -$ cold	$C -$ rain	P
0	0	0	0.3
0	0	1	0.1
0	1	0	0.15
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.1
1	1	1	0.15

What is:

A	B	$P(A, B)$
0	0	
0	1	
1	0	
1	1	

?

We can use the joint probability table to compute conditional probabilities

$A - \text{cloudy}$	$B - \text{cold}$	$C - \text{rain}$	P
0	0	0	0.3
0	0	1	0.1
0	1	0	0.15
0	1	1	0.05
1	0	0	0.05
1	0	1	0.1
1	1	0	0.1
1	1	1	0.15

What is $P(C = 1|A = 0, B = 1)$?

Recall

$$P(X_1|X_2 \ldots X_n) = \frac{P(X_1, X_2 .. X_n)}{P(X_2 .. X_n)}$$

CAN WE JUST ESTIMATE P(Y|X) IN THIS FASHION AND BE DONE?

We might not have enough data. For example consider having 100 attributes of people:

- how many rows will we have?
- how many people on earth?

CAN WE JUST ESTIMATE P(Y|X) IN THIS FASHION AND BE DONE?

We might not have enough data. For example consider having 100 attributes of people:

- how many rows will we have? $2^{100} > 10^{30}$
- how many people on earth? 10^{10}
- 99.99\% of rows will not have training examples!

SOLUTION:

1- Be smart about how to estimate probabilities from sparse data

- maximum likelihood estimates
- maximum a posteriori estimates
- Be smart about how to represent joint distributions
- Bayes networks, graphical models, conditional independencies

ESTIMATING PROBABILITIES

Assume you throw a thumbtack 10 times and it lands 3 times on its back and 7 times on its side.

What is the probability of this thumbtack falling on its side?

ESTIMATING PROBABILITIES

We model the thumbtack falling on its side as a random variable X.

Each flip of the returns a Boolean value for X, that follows a Bernouilli distribution with parameter θ:

$$X \sim \text{Bernouilli}(\theta)$$

$P(X = 1) = \theta$

$P(X = 0) = 1 - \theta$

We have observed the thumbtack falling on its side (X=1) 7 times and the thumbtack falling on its back (X=0) 3 times.

How can we estimate θ (the probability of the thumbtack falling on its side.)

MAXIMUM LIKELIHOOD ESTIMATION

The first principle we will use is the principle of Maximum Likelihood Estimation (MLE).

MLE chooses the parameter $\hat{\theta}$ that **maximize the probability of the observed data $\boldsymbol{P(\text{data}|\hat{\theta})}$**

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

$P(X = 1) = \theta$

$P(X = 0) = 1 - \theta$

We refer to observed data as $D : \{1, 0, 1, 0, 0, \ldots, 1\}$

Throwing the thumbtack produces data D with α_1 side falls (X=1) and α_0 back falls (X=0).

Throws are Independently Identically Distributed (IID).

What is $P(D|\theta)$?

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

$P(X = 1) = \theta$

$P(X = 0) = 1 - \theta$

We refer to observed data as $D : \{1, 0, 1, 0, 0, \dots, 1\}$

Throwing the thumbtack produces data D with α_1 side falls (X=1) and α_0 back falls (X=0).

Throws are Independently Identically Distributed (IID).

$$\begin{aligned} P(D|\theta) &= \theta(1-\theta)\dots\theta \\ &= \theta^{\alpha_1}(1-\theta)^{\alpha_0} \end{aligned}$$

MLE chooses the parameter $\hat{\theta}$ that **maximize the likelihood of the observed data $\boldsymbol{P(D|\hat{\theta})}$**

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

MLE chooses the parameter $\hat{\theta}_{\text{MLE}}$ that **maximize the likelihood of the observed data $\boldsymbol{P(D|\hat{\theta})}$**

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\ P(D|\theta)$$
$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\ \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

How to find $\hat{\theta}_{\text{MLE}}$?

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

MLE chooses the parameter $\hat{\theta}_{\text{MLE}}$ that **maximize the likelihood of the observed data $\boldsymbol{P(D|\hat{\theta})}$**

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta)$$

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, \ln P(D|\theta)$$

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

MLE chooses the parameter $\hat{\theta}_{\mathrm{MLE}}$ that **maximize the log likelihood of the observed data $\boldsymbol{P(D|\hat{\theta})}$**

$$\begin{aligned}\hat{\theta}_{\mathrm{MLE}} &= \underset{\theta}{\operatorname{argmax}} \ln P(D|\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \ln \left(\theta^{\alpha_1}(1-\theta)^{\alpha_0}\right)\end{aligned}$$

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

MLE chooses the parameter $\hat{\theta}_{\mathrm{MLE}}$ that **maximize the log likelihood of the observed data $\boldsymbol{P(D|\hat{\theta})}$**

$$\begin{aligned}\hat{\theta}_{\mathrm{MLE}} &= \underset{\theta}{\operatorname{argmax}} \ln P(D|\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \ln\left(\theta^{\alpha_1}(1-\theta)^{\alpha_0}\right) \ = \underset{\theta}{\operatorname{argmax}} [\alpha_1 \ln\theta + \alpha_0 \ln(1-\theta)]\end{aligned}$$

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}[\alpha_1 \ln \theta + \alpha_0 \ln(1-\theta)]$$

Take the derivative and set it to 0:

$$\frac{d}{d\theta} \ln P(D|\theta) = \frac{d}{d\theta}[\alpha_1 \ln \theta + \alpha_0 \ln(1-\theta)] = 0$$

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}[\alpha_1 \ln \theta + \alpha_0 \ln(1 - \theta)]$$

Take the derivative and set it to 0:

$$\frac{d}{d\theta} \ln P(D|\theta) = \frac{d}{d\theta}[\alpha_1 \ln \theta + \alpha_0 \ln(1 - \theta)] = 0$$

Recall

$$\frac{d}{d\theta} \ln \theta = \frac{1}{\theta}$$

$$\frac{d}{d\theta} f(g(\theta)) = g'(\theta) f'(g(\theta))$$

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} [\alpha_1 \ln \theta + \alpha_0 \ln(1-\theta)]$$

Take the derivative and set it to 0:

$$\begin{aligned} \frac{d}{d\theta} \ln P(D|\theta) &= \frac{d}{d\theta} [\alpha_1 \ln \theta + \alpha_0 \ln(1-\theta)] \\ &= \frac{\alpha_1}{\theta} - \frac{\alpha_0}{(1-\theta)} \end{aligned}$$

MAXIMUM LIKELIHOOD ESTIMATION FOR BERNOUILLI VARIABLES

SUMMARY:

Random Variable $X \sim \mathrm{Bernouilli}(\theta)$, i.e. $P(X = 1) = \theta$ and $P(X = 0) = 1 - \theta$.

We can write $P(X) = \theta^X (1 - \theta)^{1-X}$

We observe data D. We observe $X = 1$ α_1 times and $X = 0$ α_0 times.

$P(D|\theta) = \theta^{\alpha_1} (1 - \theta)^{\alpha_0}$.

We can estimate:

$$\hat{\theta}_{\mathrm{MLE}} = \underset{\theta}{\mathrm{argmax}}\ P(D|\theta) = \frac{\alpha_1}{\alpha_0 + \alpha_1}$$

PROBABILITY BOUND

In our example, we estimate $\hat{\theta}_{\text{MLE}} = 7/10$. We only threw the thumbtack 10 times.

How much can we trust this estimate?

USE HOEFFDING INEQUALITY

Given $X \sim \text{Bernouilli}(\theta)$, and D consisting of n IID samples where $X = 1$ is observed α_1 times and $X = 0$ is observed α_0 times, we have the guarantee:

$$\text{For any } \epsilon > 0, \quad P(|\hat{\theta}_{\text{MLE}} - \theta| \geq \epsilon) \leq 2\text{e}^{-2n\epsilon^2}$$

In [4]:

```
import numpy as np

epsilon = 0.1
n = 1000

p_mistake = 2*np.exp(-2*n*epsilon**2)

print("the probability of making a mistake greater than {} is less than {}".format(epsilon, p_mistake))
print("the probability of making a mistake at most {} is more than {}".format(epsilon, 1- p_mistake))
```

```
the probability of making a mistake greater than 0.1 is less than 4.122307244877101e-09
the probability of making a mistake at most 0.1 is more than 0.9999999958776927
```

HOEFFDING INEQUALITY

$$\text{For any } \epsilon > 0, \quad P(|\hat{\theta}_{\text{MLE}} - \theta| \geq \epsilon) \leq 2\mathrm{e}^{-2n\epsilon^2}$$

HIGH PROBABILITY BOUND

If we want to know how many flips we need to estimate parameter θ with error at most ϵ with a probability of $1 - \delta$.

For example: estimate parameter θ with error at most 0.01 with a probability of 0.99.

$$\text{For any } \epsilon > 0, \quad P(|\hat{\theta}_{\text{MLE}} - \theta| \geq \epsilon) \leq 2\mathrm{e}^{-2n\epsilon^2} \leq \delta$$

How may flips do we need?

HOEFFDING INEQUALITY

$$\text{For any } \epsilon > 0, \quad P(|\hat{\theta}_{\text{MLE}} - \theta| \geq \epsilon) \leq 2\mathrm{e}^{-2n\epsilon^2}$$

HIGH PROBABILITY BOUND

If we want to know how many flips we need to estimate parameter θ with error at most ϵ with a probability of $1 - \delta$.

For example: estimate parameter θ with error at most 0.05 with a probability of 0.95.

$$\text{For any } \epsilon > 0, \quad P(|\hat{\theta}_{\text{MLE}} - \theta| \geq \epsilon) \leq 2\mathrm{e}^{-2n\epsilon^2} \leq \delta$$

How may flips do we need?

$$n \geq \frac{\ln \frac{2}{\delta}}{2\epsilon^2}$$

In [5]:

```
epsilon = 0.05
delta = 0.1
minimum_n = int(np.ceil(np.log(2/delta)/(2*epsilon**2)))
print("the minimum number of flips is {}".format(minimum_n))
```

```
the minimum number of flips is 600
```

ESTIMATING PROBABILITIES

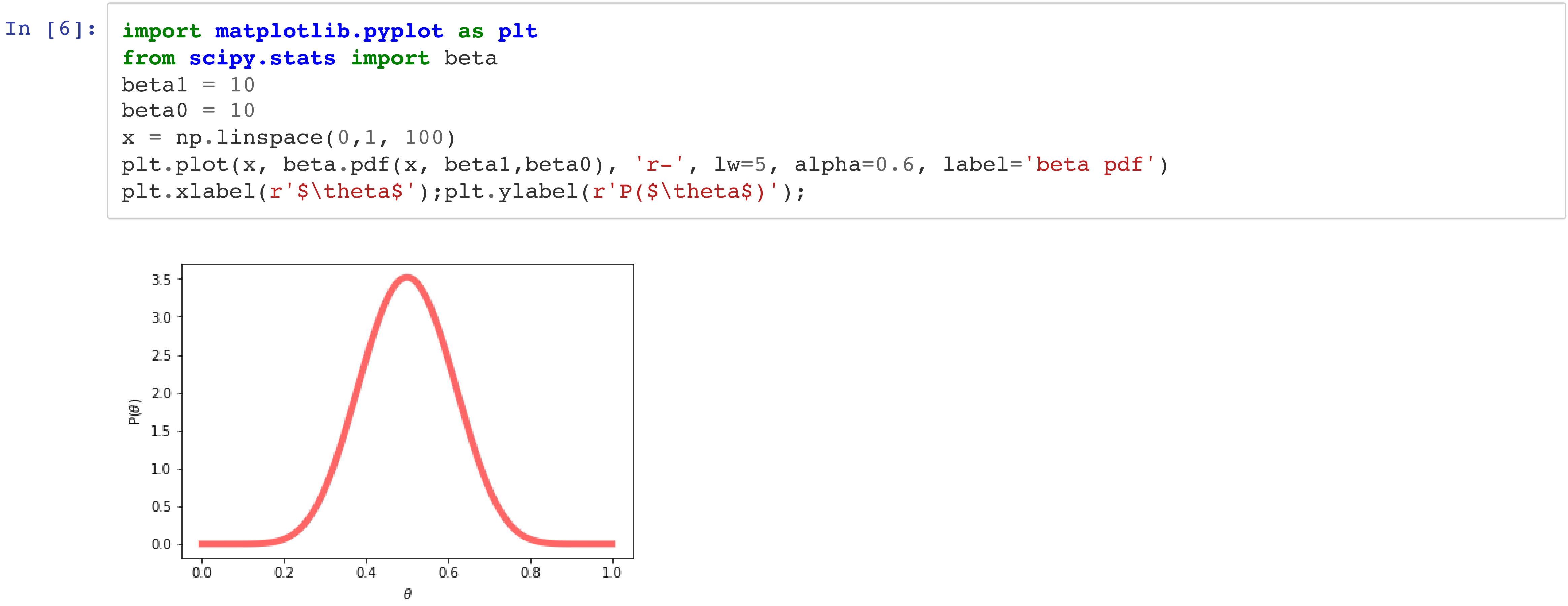
Let's go back to the original example. We threw the thumbtack 10 times, it landed 7 times on its head and 3 times on its back.

What if instead I told you that you take a coin from your pocket, you flip it lands 7 times on heads and 3 times on tails.

What is θ, the probability of heads?

ESTIMATING PROBABILITIES

You have a strong prior belief for the coin being fair. As an example of your prior distribution:



In [6]:

```
import matplotlib.pyplot as plt
from scipy.stats import beta
beta1 = 10
beta0 = 10
x = np.linspace(0,1, 100)
plt.plot(x, beta.pdf(x, beta1,beta0), 'r-', lw=5, alpha=0.6, label='beta pdf')
plt.xlabel(r'$\theta$');plt.ylabel(r'P($\theta$)');
```

BAYES RULE

Recall $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$

We use this to state:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In our context:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

BAYES RULE

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(D|\theta)$ is the **likelihood** of the data, $P(\theta)$ is the **prior** probability and $P(\theta|D)$ is the **posterior** probability of the parameter θ.

Because $P(D)$ is constant for different values of θ, it is often omitted. In that case we don't compute the exact value of the posterior but we can compare it for different values of θ:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

MAP ESTIMATION

The second principle we will use is the principle of Maximum A-posteriori Probability (MAP) Estimation.

MAP estimation chooses the parameter $\hat{\theta}_{\text{MAP}}$ that **maximize the posterior probability $\boldsymbol{P(\hat{\theta}|\text{data})}$**

$$\hat{\theta}_{\text{MAP}} = \underset{\theta}{\operatorname{argmax}}\, P(\theta|D) = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta)P(\theta)$$

Compare this with MLE, which chooses the parameter $\hat{\theta}_{\text{MLE}}$ that **maximize the likelihood of the observed data $\boldsymbol{P(D|\hat{\theta})}$**

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, P(D|\theta)$$

HOW TO CHOOSE A PRIOR?

- Uniform / uninformative prior: constant for all values of θ ==> similar to MLE.
- Can represent the prior belief
- Conjugate prior: $P(\theta)$ and $P(\theta|D)$ have same form (will see soon)

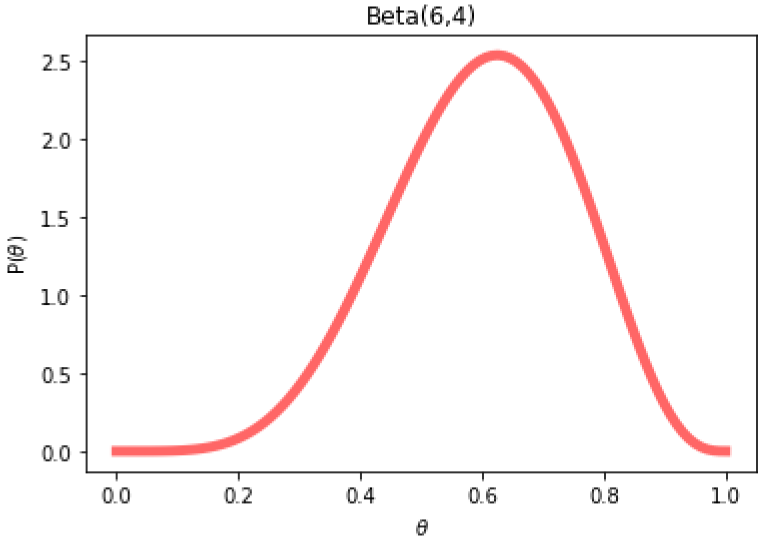
For our problem with a binomial likelihood (X is bernouilli and D binomial), the beta distribution is very useful. It has parameters β_H and β_T which can represent the prior probability.

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)}$$

In [10]:

```
beta_H = 6
beta_T = 4

plt.plot(x, beta.pdf(x, beta_H,beta_T), 'r-', lw=5, alpha=0.6, label='beta pdf')
plt.xlabel(r'$\theta$');plt.ylabel(r'P($\theta$)'), plt.title('Beta({},{})'.format(beta_H,beta_T));
```



MAP ESTIMATION

Assume you observe data D with α_H heads and α_T tails:

$$\begin{aligned} P(\theta|D) &\propto P(D|\theta)P(\theta) \\ &= \theta^{\alpha_H}(1-\theta)^{\alpha_T}\frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H,\beta_T)} \\ &\propto \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1} \end{aligned}$$

$P(\theta|D)$ is a Beta distribution with parameters $\alpha_H + \beta_H$ and $\alpha_T + \beta_T$.

Both the prior and posterior are of the same family. A binomial likelihood with a Beta prior gives a Beta posterior.

There are many other combinations of prior and likelihood functions that lead to a conjugate prior.

MAP ESTIMATION

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \underset{\theta}{\text{argmax}}\ P(\theta|D)\\ &= \underset{\theta}{\text{argmax}}\ \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}\\ &= \underset{\theta}{\text{argmax}}\ \ln[\underset{\theta}{\text{argmax}}\ \theta^{\alpha_H+\beta_H-1}(1-\theta)^{\alpha_T+\beta_T-1}]\end{aligned}$$

We already know how to solve this:

$$\underset{\theta}{\text{argmax}} = \frac{\alpha_H+\beta_H-1}{(\alpha_H+\beta_H-1)+(\alpha_T+\beta_T-1)}$$

Notice how the prior acts as additional coin flips:

- prior influences the value of θ_{MAP}
- prior can be strong or weak (high or low β_H and β_T)
- a large amount of data will reduce the effect of the prior

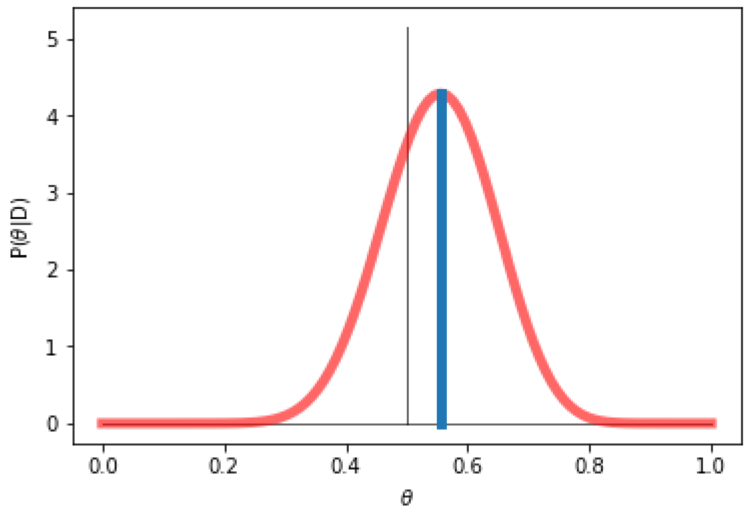
In [14]:

```
alpha_H = 6; alpha_T = 3
beta_H = 10; beta_T = 10
map_pdf = beta.pdf(x, alpha_H + beta_H, alpha_T+ beta_T)
map_value = x[np.argmax(map_pdf)]


plt.plot([0,1],[0,0], 'k', lw=1, alpha=0.6)
plt.plot([0.5,0.5],[0,np.max(map_pdf)*1.2], 'k', lw=1, alpha=0.6)
plt.plot(x, map_pdf, 'r-', lw=5, alpha=0.6, label='beta pdf')
plt.plot([map_value,map_value],[0,np.max(map_pdf)],lw=5)
plt.xlabel(r'$\theta$');plt.ylabel(r'P($\theta$|D)'), #plt.title('Beta({},{})'.format(beta_H,beta_T));
```

Out[14]:

```
(Text(0, 0.5, 'P($\\theta$|D)'),)
```



MAP ESTIMATION FOR OTHER DISTRIBUTIONS

MULTINOMIAL LIKELIHOOD AND DIRICHLET PRIOR DISTRIBUTION:

We roll a dice with 6 sides. The data D we observe consists of counts α_1, α_2 .. α_6 of observing each of the sides respectively.

The data is $\sim \text{Multinomial}(\theta_1, \theta_2, \ldots, \theta_k)$. θ is a parameter vector with k entries $(\theta_1, \theta_2, \ldots, \theta_k)$ (6 in the case of a dice).

$$P(D|\theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}.$$

We choose a Dirichlet prior $\theta \sim \text{Dirichlet}(\beta_1, \ldots \beta_k)$

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i - 1}}{B(\beta_1, \ldots \beta_k)}.$$

The posterior is also a Dirichlet: $\theta \sim \text{Dirichlet}(\alpha_1 + \beta_1, \ldots \alpha_k + \beta_k)$.

The mode of this distribution is $\hat{\theta}_{\text{MAP}}$ where the ith element is $\frac{\alpha_i + \beta_i - 1}{\sum_{k=1}^{K}(\alpha_k + \beta_k) - K}$.

For a Multinomial likelihood, the conjugate prior is Dirichlet.

CONTINUOUS VARIABLES

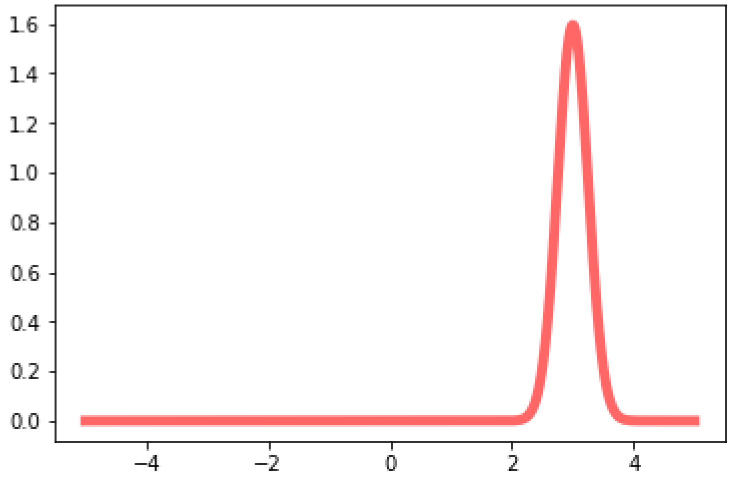
LIKELIHOOD AND PRIOR ARE NORMAL DISTRIBUTIONS.

Assume $X \sim \mathrm{N}(\mu, \sigma)$

$$p(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

```
In [19]: from scipy.stats import norm
         mu = 3
         sigma = 0.25
         x = np.linspace(-5,5, 1000)
         plt.plot(x, norm.pdf(x, mu,sigma), 'r-', lw=5, alpha=0.6, label='beta pdf')
```

```
Out[19]: [<matplotlib.lines.Line2D at 0x14308dcf8>]
```



MLE ESTIMATION FOR MEAN OF GAUSSIAN

You observe n IID samples $D = \{x_1, x_2, \ldots x_n\}$.

$$\begin{aligned} P(D|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \ \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2}\Big) \\ &= \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^n \ \exp\Big(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\Big) \end{aligned}$$

We want to find μ_{MLE}:

MLE ESTIMATION FOR MEAN OF GAUSSIAN

You observe n IID samples $D = \{x_1, x_2, \ldots n_n\}$.

$$\begin{aligned} P(D|\mu,\sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \, \exp\Big(-\frac{(x_i - \mu)^2}{2\sigma^2}\Big) \\ &= \Big(\frac{1}{\sigma\sqrt{2\pi}}\Big)^n \, \exp\Big(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\Big) \end{aligned}$$

We want to find μ_{MLE}:

$$\frac{d}{d\mu}\ln P(D|\mu,\sigma) = \frac{d}{d\mu}\ln\exp\Big(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\Big) = \frac{d}{d\mu} - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2}$$

Setting to 0 and solving gives

$$\mu_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n}$$

MAP ESTIMATION FOR MEAN OF GAUSSIAN

Using a normal prior for $\mu \sim N(\mu_0, \sigma_0)$ results in a normal posterior as well.

(Can do it on your own).

WHAT YOU SHOULD KNOW:

- revise concepts of joint / marginal / conditional probability
- bayes rule
- MLE estimation principles
- MAP estimation principles
- Bernouilli, Binomial, Multinomial, Beta, Dirichlet and Normal distributions