Recall: Bayes Optimal Classifier

If provided w/ $\mathbb{P}(X,Y)$ we can predict the most likely label for $\vec{x}$, formally $\underset{y}{\operatorname{argmax}}\ \mathbb{P}(y|\vec{x})$

It's therefore worth considering if we can estimate $\mathbb{P}(X,Y)$ directly from the training data.

If this is possible we could then use Bayes Optimal Classifier in practice on our estimate of $\mathbb{P}(X,Y)$

Much of supervised learning falls into two categories

1. Generative: estimating $\mathbb{P}(X|Y)$ $\quad \mathbb{P}(X,Y) = \mathbb{P}(X|Y)\,\mathbb{P}(Y)$
2. Discriminative: estimating $\mathbb{P}(Y|X)$ $\quad \mathbb{P}(X,Y) = \mathbb{P}(Y|X)\,\mathbb{P}(X)$

Example: Coin Toss

Toss a coin 10-times

$$D = \{\underset{x_1}{H}, T, T, H, H, H, T, T, T, \underset{x_n}{T}\}$$

Based on these samples, how would you estimate $\mathbb{P}(H)$?

Intuitively,

$$\mathbb{P}(H) = \frac{n_H}{n_H + n_T} = \frac{4}{10} = 0.4$$

Can we derive this more formally?

Maximum Likelihood Estimation

Proceed in two steps

1. Make an explicit modeling assumption about what type of distribution your data was sampled from

2. Set the parameters of this distribution so that the data you observed is as likely as possible.

In coin example, we assume the observed outcomes come from a binomial distribution.

Let $n \rightarrow$ number of coin tosses

$\theta \rightarrow$ probability of coin being heads $(P(H) = \theta)$

Formally, binomial distribution is defined by

$$P(D|\theta) = \binom{n_H + n_T}{n_H} \theta^{n_H} (1-\theta)^{n_T}$$

and it computes the probability that we observe exactly n_H heads and n_T tails if a coin was tossed $n = n_H + n_T$ times and its probability of coming up heads is θ

MLE Principle: Find $\hat{\theta}$ to maximize likelihood of the data $\mathbb{P}(D;\theta)$

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} \mathbb{P}(D;\theta)$$

This is a maximization problem!

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} \mathbb{P}(D;\theta)$$

$$= \operatorname{argmax}_{\theta} \binom{n_H + n_T}{n_H} \theta^{n_H} (1-\theta)^{n_T}$$

Take log-likelihood

$$\log(\hat{\theta}_{MLE}) = \operatorname{argmax}_{\theta} \log\binom{n_H + n_T}{n_H} + n_H \log(\theta) + n_T \log(1-\theta)$$

Take derivative, set to zero $(d/d\theta)$

$$\frac{n_H}{\theta} - \frac{n_T}{1-\theta} = 0$$

θ that solves this is $\hat{\theta}_{MLE}$

$$\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T}$$ } sanity check; $\theta \in [0,1]$

verify this is a max by taking second derivative and assuring it's always negative!

So,

- MLE gives the explanation of the data you observed
- If n is large and your model is correct ($\mathcal{H}$ includes true model), then MLE finds **true** parameters
- **BUT**, MLE overfits the data if n is small

If you don't have the correct model (and n is small) then MLE can be terribly wrong!

Example: Coin Toss with Prior Knowledge

Assume you have a hunch θ is close to 0.5. But your sample size is small, so you don't trust your estimate.

Simple fix: Add m imaginary throws that would result in θ' (e.g. $\theta = 0.5$). Add m H and m T to data.

$$\hat{\theta} = \frac{n_H + m}{n_H + n_T + 2m}$$

insignificant change for large n.

Incorporates "prior" belief about what θ should be for small n.

Can we derive this?

The Bayesian Way

Model θ as a **random variable**, drawn from a distribution $P(\theta)$

Note: θ is **NOT** a random variable associated w/ an event in a sample space

In Bayesian statistics, this is allowed and you can specify a prior belief $P(\theta)$ defining what values you believe θ is likely to take on.

Observe

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

where

- $P(\theta)$ is the **prior** distribution over θ, before data is observed
- $P(D|\theta)$ is the likelihood of the data given the parameter θ
- $P(\theta|D)$ is the posterior distribution over θ **after** we have observed the data

A natural choice for the prior $\mathbb{P}(\theta)$ is the **Beta distribution**

$$\mathbb{P}(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

where

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

Multivariate generalization is called Dirichlet distribution

is the normalization constant.

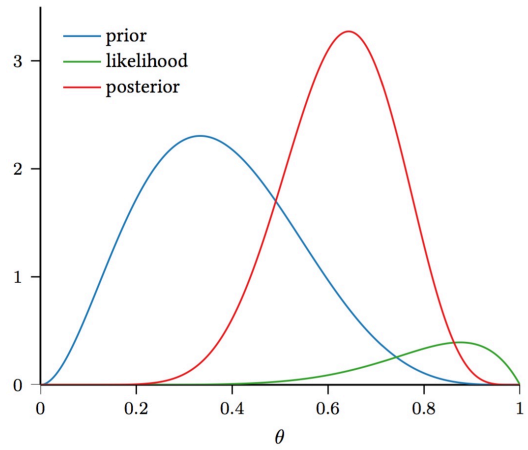
Why use Beta distribution?

- It models probabilities $(\theta \in [0,1])$
- It is in the same distributional family as the binomial distribution (conjugate prior)

i.e ← which make math work out nicely:

$$\mathbb{P}(\theta|D) \propto \mathbb{P}(D|\theta)\mathbb{P}(\theta) \propto \theta^{n_H+\alpha-1}(1-\theta)^{n_T+\beta-1}$$

So we have a distribution over θ, how do we get an estimate?



Maximum a Posteriori Probability Estimation (MAP)

(Frequentist Approach)

MAP Principle: Find $\hat{\theta}$ that maximizes the posterior $\mathbb{P}(\theta|D)$

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

$$\underset{\theta}{\text{argmax}}\ \log(\mathbb{P}(D|\theta)) + \log(\mathbb{P}(\theta))$$

For our coin flipping scenario, we get

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

$$= \underset{\theta}{\text{argmax}}\ \frac{\mathbb{P}(D|\theta)\,\mathbb{P}(\theta)}{\mathbb{P}(D)}$$

$$= \underset{\theta}{\text{argmax}}\ \log(\mathbb{P}(D|\theta)) + \log(\mathbb{P}(\theta))$$

$$= \underset{\theta}{\text{argmax}}\ n_H \log(\theta) + n_T \log(1-\theta) + (\alpha-1)\log(\theta) + (\beta-1)\log(1-\theta)$$

$$= \underset{\theta}{\text{argmax}}\ (n_H + \alpha - 1)\log(\theta) + (n_T + \beta - 1)\log(1-\theta)$$

$$\frac{d}{d\theta} \downarrow \qquad \frac{n_H + \alpha - 1}{\theta} - \frac{n_T + \beta - 1}{1-\theta} = 0$$

$$\hat{\theta}_{MAP} = \frac{n_H + \alpha - 1}{n_H + n_T + \alpha + \beta - 2}$$

A few comments (on MAP):

(i) MAP estimate is identical to MLE w/ $(\alpha-1)$ hallucinated heads and $(\beta-1)$ hallucinated tails

(ii) As $n \to \infty$, $\hat{\theta}_{MAP} \to \hat{\theta}_{MLE}$ as $\alpha-1$ and $\beta-1$ become irrelevant compared to large n_H, n_T.

(iii) MAP is a great estimator if an accurate prior belief is available

(iv) If n small, MAP can be VERY wrong if prior belief is wrong

"True" Bayesian Approach

MAP is one way to get an estimator. But, there's much more info in $\mathbb{P}(\theta|D)$!

A true Bayesian approach is to use the posterior predictive distribution directly to make predictions about the label Y of a test sample with features X:

$$\mathbb{P}(Y|D,X) = \int_\theta \mathbb{P}(Y,\theta|D,X)\,d\theta = \int_\theta \mathbb{P}(Y|\theta,D,X)\,\mathbb{P}(\theta|D)\,d\theta$$

↑ prediction take into consideration ALL possible models

↗ unfortunately intractable in closed form, but techniques such as Monte Carlo approximate the distribution

Our previous coin example is actually an exception to this

$$\mathbb{P}(\text{heads} \mid D) = \int_\theta \mathbb{P}(\text{heads}, \theta \mid D)\, d\theta$$

← Chain rule! $\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid B, C)\,\mathbb{P}(B \mid C)$

$$= \int_\theta \mathbb{P}(\text{heads} \mid \theta, D)\,\mathbb{P}(\theta \mid D)\, d\theta$$

$$= \int_\theta \theta\, \mathbb{P}(\theta \mid D)\, d\theta$$

$$= \mathbb{E}[\theta \mid D]$$

$$= \frac{n_H + \alpha}{n_H + n_T + \alpha + \beta}$$

Here, we use the fact that we defined

$$\mathbb{P}(\text{heads} \mid D, \theta) = \mathbb{P}(\text{heads} \mid \theta) = \theta$$

and this only holds b/c we assumed our data is drawn from a binomial distribution - doesn't hold in general

Machine Learning and Estimation

In supervised learning you are provided w/ training data D. You use this data to train a model, represented by its parameters θ. With this model you wanna make predictions on a test point x_t.

MLE Prediction: $\mathbb{P}(y|x_t;\theta)$ Learning: $\theta = \underset{\theta}{\text{argmax}}\, \mathbb{P}(D;\theta)$.
Here θ is purely a model parameter

MAP Prediction: $\mathbb{P}(y|x_t,\theta)$ Learning: $\theta = \underset{\theta}{\text{argmax}}\, \mathbb{P}(\theta|D)$
where $\mathbb{P}(\theta|D) \propto \mathbb{P}(D|\theta)\mathbb{P}(\theta)$.
Here θ is a random variable

"True Bayesian" Prediction: $\mathbb{P}(y|x_t,D) = \int_\theta \mathbb{P}(y|\theta)\mathbb{P}(\theta|D)\,d\theta$
Here θ is integrated out! Our prediction takes all possible models into account.

Difference b/t MLE and MAP is subtle!

MLE we maximize $\log(\mathbb{P}(D;\theta))$
MAP we maximize $\log(\mathbb{P}(D|\theta)) + \log(\mathbb{P}(\theta))$

independent of data and penalizes if θ deviates from what we believe is reasonable