Probability for Machine Learning – Naïve Bayes, MLE, MAP

Aditya Arun

IIIT Hyderabad

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

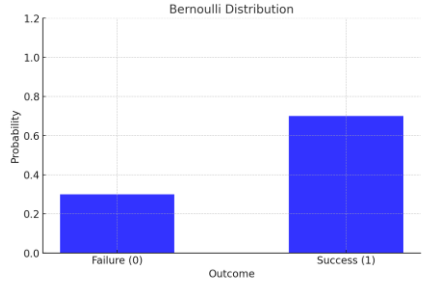
Probability Fundamentals Recap

Recap – Data as Distribution

- Data as distributions
  - Bernoulli

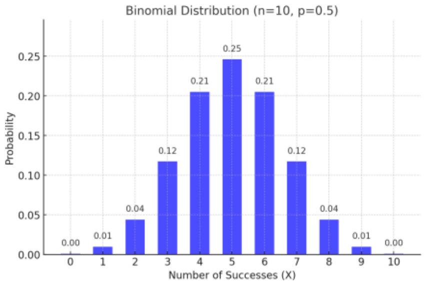
$$P(X = x) = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0. \end{cases}$$

$$P(X = x) = p^x (1 - p)^{1-x}, \quad x \in \{0, 1\}$$



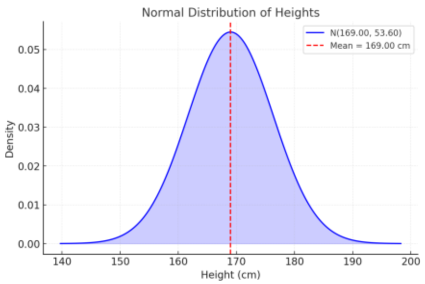
  - Binomial

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$



  - Gaussian

$$P(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Recap – Probability Basics

- Basic Probability Concepts
  - expectation $$\mathbb{E}[X] = \sum_x x \cdot P(X = x) = \int_{-\infty}^{\infty} x \cdot P(X) dx$$
  - variance $$Var(X) = \mathbb{E}[(X - \mu)^2]$$
  - joint probability $$P(X, Y)$$
  - sum rule or marginalization $$P(X) = \sum_y P(X, Y) = \int_y P(X, Y) dy$$
  - conditional probability $$P(Y|X) = \frac{P(X, Y)}{P(X)}$$
  - product rule $$P(X, Y) = P(X|Y) P(Y)$$
  - law of total probability $$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i) = \int_{-\infty}^{\infty} P(A|B_i) P(B_i) db$$
  - Bayes theorem $$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Bayesian Classifier and Naïve Bayes

Learning Classifiers as Bayes Rule

- Supervised Learning Problem: $f : X \to Y$, or equivalently $P(Y|X)$
  - Y is a Boolean value random variable, $X = \langle X_1, X_2, \ldots, X_n \rangle$
- Applying Bayes Rule,

$$P(Y = y_i | X = x_k) = \frac{P(X = x_k | Y = y_i) P(Y = y_i)}{\sum_j P(X = x_k | Y = y_j) P(Y = y_j)}$$

- A simple objective
  - Decide y_1, if $P(Y = y_1 | X = x_k) > P(Y = y_2 | X = x_k)$
  - Decide y_2, if $P(Y = y_1 | X = x_k) < P(Y = y_2 | X = x_k)$
  - Or, $Y \leftarrow \arg\max P(Y|X)$
- Is it practical to compute?
  - Think how much data is required to estimate each distribution

Unbiased Learning of Bayes Classifier is Infeasible

Total Number of Parameters

- Consider the distribution $P(X = x_k | Y = y_i)$
  - Number of parameters in this unknown distribution?

$$\theta_{i,j} \equiv P(X = x_i | Y = y_j)$$

  - i is indexed on 2^n possible values, one for each possible values of X
  - j is indexed on 2 possible values.

Reducing Number of Parameters Due to Constraints

- For a fixed j, the sum over i in $\theta_{i,j}$ must be one. $\sum_i \theta_{i,j} = 1$
  - This constraint removes one degree of freedom for each j, meaning that for each y_j, we need to estimate only $2^n - 1$ independent parameters.
  - For two values of Y, the total number of independent parameters is $2(2^n - 1) = 2^{n+1} - 2$.
  - As n grows large, the number of parameters to estimate grows exponentially $O(2^{n+1})$.
    - **For $n = 30$, the total number of parameters is 2 billion!**

Naïve Bayes Algorithm

- Assumes conditional independence when modeling $P(X|Y)$
  - Reduces complexity from $O(2^{n+1})$ to $O(2n)$

Conditional Independence

- Given three sets of random variable X, Y, and Z.
- X is conditionally independent of Y given Z, when:

$$(\forall i, j, k) P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

Derivation of Naïve Bayes Algorithm

- Goal, learn $P(X|Y)$, where $X = \langle X_1, X_2, \ldots, X_n \rangle$,
- Naïve Bayes algorithm assumes independence between $X_1, X_2, \ldots, X_n$ given Y.
- A simple case:

$$\begin{aligned} P(X|Y) &= P(X_1, X_2|Y) \\ &= P(X_1|X_2, Y)P(X_2|Y) \\ &= P(X_1|Y)P(X_2|Y) \end{aligned}$$

- Generally:

$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Derivation of Naïve Bayes algorithm

- Applying Bayes Rule,

$$P(Y = y_k|X_1, \ldots, X_n) = \frac{P(Y = y_k)P(X_1, \ldots, X_n|Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \ldots, X_n|Y = y_j)}$$

- Assuming Conditional Independence,

$$P(Y = y_k|X_1, \ldots, X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

- For most probable value of Y, we have Naïve Bayes classification rule,

$$Y \leftarrow \arg\max_{y_k} \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

- This simplifies to:

$$\boxed{Y \leftarrow \arg\max_{y_k} P(Y = y_k)\prod_i P(X_i|Y = y_k)}$$

Estimating Probabilities

Joint Probability Distributions

The key to probabilistic models is defining random variables and their joint distribution

- Consider the following table with joint distribution defined over 3 variables:

Gender	HoursWorked	Wealth	probability
female	< 40.5	poor	0.2531
female	< 40.5	rich	0.0246
female	≥ 40.5	poor	0.0422
female	≥ 40.5	rich	0.0116
male	< 40.5	poor	0.3313
male	< 40.5	rich	0.0972
male	≥ 40.5	poor	0.1341
male	≥ 40.5	rich	0.1059

Importance of Joint Probability Distribution

- Joint Probability Distribution is central to probabilistic inference, we can answer any possible probabilistic question that can be asked about these variables
  - The joint probability allows computing any conditional or joint probability over any subset of variables.
- Computing Marginal Probabilities
  - $P(Gender = Male) = 0.6685$
  - $P(Wealth = rich) = 0.2393$
- Computing joint probabilities over subset of variables
  - $P(Wealth = rich \wedge Gender = female) = 0.0362$
- Computing Conditional Probabilities
  - $P(Wealth = rich|Gender = Female) = \frac{0.0362}{0.3315} = 0.1092$

Gender	HoursWorked	Wealth	probability
female	< 40.5	poor	0.2531
female	< 40.5	rich	0.0246
female	≥ 40.5	poor	0.0422
female	≥ 40.5	rich	0.0116
male	< 40.5	poor	0.3313
male	< 40.5	rich	0.0972
male	≥ 40.5	poor	0.1341
male	≥ 40.5	rich	0.1059

Learning Joint Probability Distribution

- Learning joint distributions from observed training data involves estimating probabilities for joint assignments in a table.
- With a large dataset (e.g., a million people), probabilities can be estimated by calculating the fraction of entries that satisfy each joint assignment.
- Reliable probability estimates are possible if each row has a large number of entries.
- Learning joint distributions can be difficult when the dataset is very large due to the exponential growth in table size as the number of features increases.
  - For example, with 100 boolean features, the table would have 2^{100} rows (more than 10^{30}), making it infeasible to obtain sufficient training data for each row.
- Effective Probability Learning requires:
  - Smart estimation of probability parameters from data
  - Efficient representation of joint probability distributions

Estimating Probabilities

- Consider we have a coin X.
  - Flipping it may turn up heads ($X = 1$) or tails ($X = 0$)

- The learning task is to estimate the probability that it will turn up heads $P(X = 1)$
- θ are the parameters of this “true” but unkown distribution (actual ground truth)
- $\hat{\theta}$ are the learned estimate of this true θ
- You gather training data by flipping coins n times
  - You observe α_1 heads and α_0 tails
  - $n = \alpha_1 + \alpha_0$

Estimating Probabilities – Algorithm 1

- What is the most intuitive approach of estimating $\theta = P(X = 1)$?

$$\hat{\theta} = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

- What happens if our training data has 50 coin flip trials with 24 heads and 26 tails?
  - $X = \langle H, T, T, H, \dots, T \rangle$
- What happens if our training data has 3 coin flip trials with 1 heads and 2 tails?
  - Unreliable estimates!

Estimating Probabilities – Algorithm 2

- Add imaginary coin flips – this reflects our prior
  - γ_1 imaginary heads, and γ_0 imaginary tails

$$\hat{\theta} = \frac{\alpha_1 + \gamma_1}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$$

- Algorithm 2, like Algorithm 1, estimates the proportion of heads while incorporating priors via imaginary flips.
- Advantages:
  - Easy to incorporate prior assumptions about the value of θ by adjusting the ratio of γ_1 and γ_2
  - Easy to incorporate degree of uncertainty about our prior knowledge by adjusting the total volume of the imaginary flips
    - For $\theta = 0.7$, what happens if we have $\gamma_1 = 700$ and $\gamma_0 = 300$ vs $\gamma_1 = 7$ and $\gamma_0 = 3$?
  - Algorithm 1 can be recovered by applying $\gamma_1 = \gamma_0 = 0$
  - Asymptotically, As observed data approaches infinity, imaginary data's effect vanishes.

Formulating probabilistic objective in ML problems

- We have data $\mathcal{D}$ and we assume it is sampled from some distribution
- How do we figure out the **parameters that best “fit” that distribution**?

Revisiting Bayes Theorem

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

- θ: model parameters
- $\mathcal{D}$: observed data
- $P(\theta|\mathcal{D})$: Posterior (probability of the parameters given the data).
- $P(\mathcal{D}|\theta)$: Likelihood (probability of data given the parameters)
- $P(\theta)$: Prior belief about parameters

Maximum Likelihood Estimate (MLE)

- Estimate the parameters θ, based on the principle that if we observe training data $\mathcal{D}$, we should choose the value of θ that makes $\mathcal{D}$ most probable

$$\theta^{\text{MLE}} = \arg\max_{\theta} P(\mathcal{D}|\theta)$$

- Intuition: we are more likely to observe data $\mathcal{D}$ if we were in world were appearance of this data is highly probable.

MLE for Coin Flip Example

- Let X be a random variable that can take either value 0 or 1
- Let $\theta = P(X = 1)$ refer to the true but unkown probability distribution
- We observe $X = 1$ a total of α_1 times, and $X = 0$ a total of α_0 times
- Assume all trials are i.i.d.

- Maximum Likelihood Estimate involves choosing θ to maximize $P(\mathcal{D}|\theta)$
  - Equivalently, $P(\alpha_1, \alpha_2|\theta)$

$$P(\mathcal{D} = \langle \alpha_1, \alpha_0 \rangle | \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

MLE for Coin Flip Example

- Data Likelihood or data likelihood function

$$L(\theta) = P(\mathcal{D} = \langle \alpha_1, \alpha_0 \rangle | \theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

- Now, we need to determine the value of θ that maximizes the data likelihood function
- Maximizing $P(\mathcal{D}|\theta)$ is equivalent to maximizing its logarithm

$$\arg\max_\theta P(\mathcal{D}|\theta) = \arg\max_\theta \ln P(\mathcal{D}|\theta)$$

- To maximize $\ln P(\mathcal{D}|\theta)$ (and thus $P(\mathcal{D}|\theta)$), take its derivative with respect to θ.
- Solve for θ where the derivative equals zero.
- Since $\ln P(\mathcal{D}|\theta)$ is concave in θ, this point is the maximum.

MLE for Coin Flip Example

- Calculate the derivative of log likelihood function

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

MLE for Coin Flip Example

- Set the derivative to zero and solve for θ

$$
\begin{aligned}
0 &= \alpha_1 \frac{1}{\theta} - \alpha_0 \frac{1}{1-\theta} \\
\alpha_0 \frac{1}{1-\theta} &= \alpha_1 \frac{1}{\theta} \\
\alpha_0 \theta &= \alpha_1 (1-\theta) \\
(\alpha_1 + \alpha_0)\theta &= \alpha_1 \\
\theta &= \frac{\alpha_1}{\alpha_1 + \alpha_0}
\end{aligned}
$$

- MLE for Coin Flip

$$\hat{\theta}^{\mathrm{MLE}} = \arg\max_{\theta} P(\mathcal{D}|\theta) = \arg\max_{\theta} \ln P(\mathcal{D}|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Maximum a Posteriori Estimate (MAP)

- Estimate the parameters θ based on the principle that we should choose the value of θ that is most probable, given the observed data $\mathcal{D}$ and our prior assumption summarized by $P(\theta)$

$$
\begin{aligned}
\hat{\theta}^{\mathrm{MAP}} &= \arg\max_{\theta} P(\theta|\mathcal{D}) \\
&= \arg\max_{\theta} \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} \\
&= \arg\max_{\theta} P(\mathcal{D}|\theta)P(\theta)
\end{aligned}
$$

- Compared to MLE, MAP has an extra prior term $P(\theta)$
  - Prior acts as a regularization term for the above objective

MAP for Coin Toss Example

- Specify a prior $P(\theta)$
  - In the coin flip example, i.i.d. trials of coins draws a Bernoulli random variable, we will use its conjugate distribution, beta distribution, to model the prior

$$P(\theta) = Beta(\beta_0, \beta_1) = \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_0 - 1}}{B(\beta_0, \beta_1)}$$

  - The denominator is a normalization term, independent of θ

MAP for Coin Toss Example

- Substituting to the MAP objective we get:

$$\begin{aligned}
\hat{\theta}^{\mathrm{MAP}} &= \arg\max_{\theta} P(\mathcal{D}|\theta)P(\theta) \\
&= \arg\max_{\theta} \theta^{\alpha_1}(1-\theta)^{\alpha_0} \frac{\theta^{\beta_1-1}(1-\theta)^{\beta_0-1}}{B(\beta_0,\beta_1)} \\
&= \arg\max_{\theta} \frac{\theta^{\alpha_1+\beta_1-1}(1-\theta)^{\alpha_0+\beta_0-1}}{B(\beta_0,\beta_1)} \\
&= \arg\max_{\theta} \theta^{\alpha_1+\beta_1-1}(1-\theta)^{\alpha_0+\beta_0-1}
\end{aligned}$$

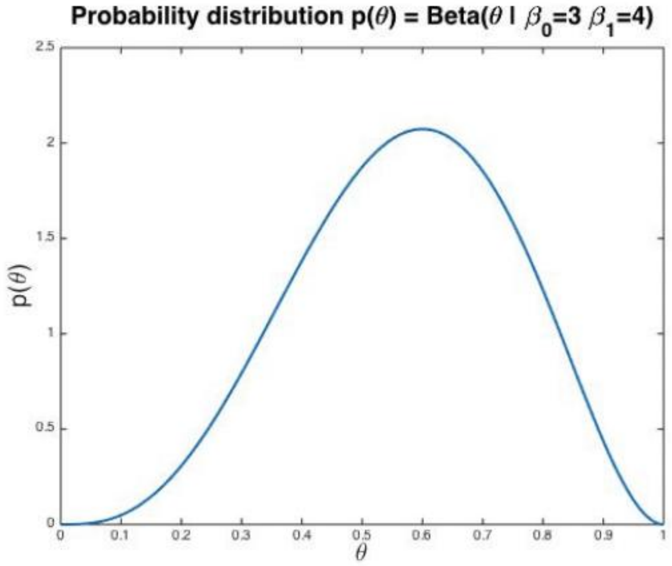
- Like MLE, solve for θ

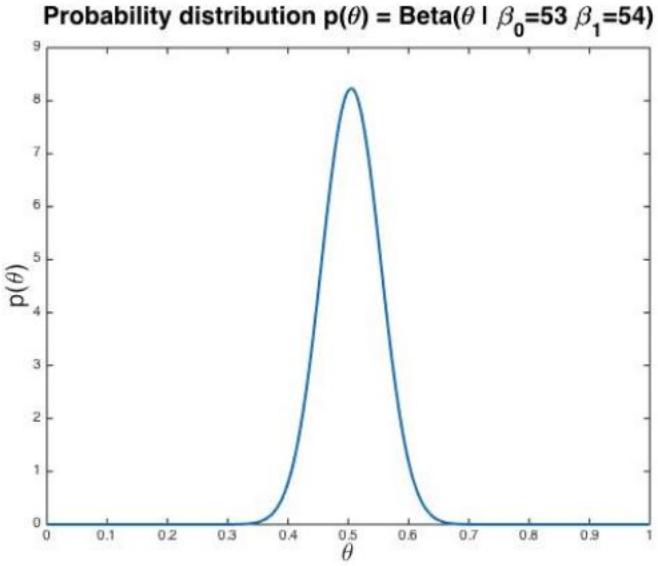
$$\hat{\theta}^{\mathrm{MAP}} = \arg\max_{\theta} P(\mathcal{D}|\theta)P(\theta) = \frac{(\alpha_1+\beta_1-1)}{(\alpha_1+\beta_1-1)+(\alpha_0+\beta_0-1)}$$

  - Draw parallels with Algorithm 2, $\gamma_1 = \beta_1 - 1$ and $\gamma_0 = \beta_0 - 1$

MAP Priors and Posteriors

- Why did we choose the $Beta(\beta_0, \beta_1)$ probability distributions to define our prior?
  - Beta distribution has a functional form that is same as data likelihood term in our problem
    - Simplifies the computation
  - The parameters β_0, β_1 play the same role as γ_0, γ_1 in Algorithm 2





- What is Conjugate Prior and Conjugacy?[1]

1. Probability Distribution and Conjugate Priors – Simon Price

MLE vs MAP

Aspect	MLE	MAP
Objective	Maximizes the likelihood $P(\mathcal{D} \mid \theta)$	Maximizes the posterior $P(\theta \mid \mathcal{D})$
Incorporates Prior?	No	Yes
Formula	$\theta_{\text{MLE}} = \arg\max_\theta P(\mathcal{D} \mid \theta)$	$\theta_{\text{MAP}} = \arg\max_\theta P(\mathcal{D} \mid \theta) \cdot P(\theta)$
Interpretation	Only considers the fit to the observed data.	Considers both data fit and prior knowledge about θ.
Objective in ML	Corresponds to minimizing only the loss term (data-fit).	Corresponds to minimizing loss + regularization.
Prior Assumption	Assumes a uniform prior (or no prior).	Allows for specific priors (e.g., Gaussian, Laplace).
Example in ML	Logistic regression without regularization.	Ridge regression (Gaussian prior), Lasso (Laplace prior).
When to Use?	When no prior knowledge is available or justified.	When prior knowledge or beliefs about θ exist.