ML
MAP
NAIVE BAYES

DENSITY ESTIMATION W. A GAUSSIAN

\[ P(\bar{x} | \theta) = \frac{1}{(2\pi)^{\frac{m}{2}}} e^{-\frac{(\theta - \bar{x})^2}{2}} \]

POINT PARAMETER

UNIT VARIANCE GAUSSIAN
\( \theta \) IS MEAN PARAMETER
DATA: \( D = (x_1, \ldots, x_T) \) \( T \) POINTS \( \in \mathbb{R}^n \)

WHICH PARAMETER EXPLAINS DATA BEST?

\( \Theta^* \) \( \text{MAXIMUM LIKELIHOOD} \)

\[ \Theta^* = \arg \max_{\theta} P(D | \theta) \]

\( \Theta^* \) \( \text{INDEPENDENT EXAMPLES} \)

\[ = \arg \max_{\theta} \prod_{t} P(x_t | \theta) \]

\[ = \arg \min_{\theta} -\log \prod_{t} P(x_t | \theta) \]

\[ = \arg \min_{\theta} \sum_{t} -\log P(x_t, \theta) \]

\( \Theta^* \) \( \text{NEGATIVE LOG LIKELIHOOD} \)

\[ = \arg \min_{\theta} \sum_{t} (\theta - x_t)^2 / 2 \]

\( \Theta^* \) \( \text{TOTAL SQUARE LOSS TO MEAN} \)

\[ \frac{\partial}{\partial \theta} \text{TOTAL LOSS} = \sum_{t} (\theta - x_t) = 0 \]

\[ t \theta^* - \sum_{t} x_t = 0 \]

\[ \theta^* = \frac{1}{T} \sum_{t=1}^{T} x_t \]

\( \theta^* \) \( \text{MEAN} \)

\( \text{EMPIRICAL MEAN} \)
OFTEN ADDITIONAL SHRINKAGE LEADS TO BETTER BOUNDS

\[ \theta^* = \frac{\sum_{t=1}^{T} x_t + \theta}{T+1} \]

ADDITIONAL DATA PT. AT \( \theta \)

\[ \text{one more} \]

\[ = \arg \min_{\theta} \left( \frac{1}{2} (G - C)^2 + \sum_{t=1}^{T} \frac{(G - x_t)^2}{2} \right) \]

LOSS WRIT ADDITIONAL PT. \( \theta \)

TOTAL LOSS ON ALL OTHER PTS

\[ = \arg \max_{\theta} \pi \prod_{t} e^{-\frac{(G - x_t)^2}{2}} \]

\[ \sim p(\theta | D) \sim p(D | \theta) \]

\[ \uparrow \quad \uparrow \]

PRIOR DATA LIKELIHOOD

\[ \sim p(\theta | D) = \frac{p(\theta) p(D | \theta)}{p(D)} \]

ABOVE ALSO CALLED MAP ESTIMATOR

\[ \text{T MAXIMUM A POSTERIORI} \]
Previous method for deriving updates

\[ w^* = \text{argmin} \quad \Delta_F (w, w_0) + \sum_t \Delta_H (w, x_t, y_t) \]

\[ \text{log prior wrt additional examples} \]

\[ \text{log likelihood of data} \]

MAP estimation
\[ P(x|\theta) \sim e^{\theta x - G(\theta)} \]

**Loss:** \[-\log P(x|\theta)\]

\[ = G(\theta) - \theta x \]

\[ \theta_{ML} = \arg \max_{\theta} \prod_x e^{\theta x - G(\theta)} \]

\[ = \arg \min_{\theta} \sum_x G(\theta) - \theta \bar{x}_t \]

\[ = \arg \min_{\theta} T G(\theta) - \theta \sum_{t} x_t = 0 \]

\[ \frac{\partial}{\partial \theta} T G(\theta) - \theta \sum_{t} x_t = T g(\theta) - \sum_{t} x_t = 0 \]

\[ g(\theta) = \frac{\sum_{t} x_t}{T} \]

\( \theta \): **natural parameter**

\( g(\theta) = \mu \): **expectation parameter**

**In case of Gaussians**

\( g \) is identity function
BERNOULLI

\( x_t \in \{0, 1\} \)  COIN FLIPS

\[
P(x | \theta) = \exp \left( \theta x - \ln (1 + e^\theta) \right)
\]

\[
\theta(\theta) = \frac{e^\theta}{1 + e^\theta}
\]

\[D = (x_1, \ldots, x_T) \]  \( T \) FLIPS

ML:
\[
g(\theta_{ML}) = \frac{e^\theta_{ML}}{1 + e^\theta_{ML}} = \frac{\sum x_t}{T}
\]

\[
g(\theta) = \frac{e^\theta}{1 + e^\theta} = \mu \quad \text{EXPECTATION PARAM.}
\]

\[
\theta = \ln \frac{\mu}{1 - \mu}
\]

LIKELIHOOD OF BERNOLLI 1 TO EXPECTATION PARAMETER

\[
P(x | \mu) = \exp \left( \ln \frac{\mu}{1 - \mu} x - \ln \left( 1 + e^{\ln \frac{\mu}{1 - \mu}} \right) \right)
\]

\[
= \exp \left( \ln \frac{\mu}{1 - \mu} x - \ln \left( 1 + \frac{\mu}{1 - \mu} \right) \right)
\]

\[
= \exp \left( \ln \frac{\mu}{1 - \mu} x - \ln \left( 1 - \mu \right) \right)
\]

\[
= \exp \left( x \ln \mu + (1 - x) \ln (1 - \mu) \right)
\]

\[
= \mu^x (1 - \mu)^{1-x} = \begin{cases} 
\mu^x & \text{when } x = 1 \\
(1 - \mu)^{1-x} & \text{when } x = 0 
\end{cases}
\]
\[ \text{LOSS: } -\log \ P(x|\mu) = -x \ln \mu - (1-x) \ln (1-\mu) \]

\[ x \in [0,1] \]

\[ x \in [0,1] \]

\[ -\log P(x|\mu) = x \ln \frac{x}{\mu} + (1-x) \ln \frac{1-x}{1-\mu} \]

**Binary Relative Entropy Between**

\[(x, 1-x) \text{ and } (\mu, 1-\mu)\]

\[ M_{\text{ML}} = \frac{\sum_{t=1}^{T} x_t}{T} \]

\[ M_{\text{KT}} = \frac{\sum_{t=1}^{T} x_t + \frac{1}{2}}{T+1} \]

\[ \text{Krichevsky-Trofimov Estimator} \]

\[ M_{\text{LAPL}} = \frac{\sum_{t=1}^{T} x_t + 1}{T+2} \]

\[ \text{Laplace Estimator} \]

\[ \text{More stable} \]

\[ \text{Has MAP interpretation w.r.t. some prior} \]

\[ \text{Additional 2/3 example at } x = 0 \quad x = 1 \]

\[ \text{Additional example at } x = 0 \quad \text{and } x = 1 \]
MULTINOMIAL

\[ x_t = (0, 0, 1, 0, 0, 0) \]

\[ k \text{ OUTCOMES ENCODED AS } k \text{ UNIT VECTORS} \]

MORE CONVENIENT TO WRITE IT TO

\[ (\mu_1, \ldots, \mu_k), \quad \mu_i \geq 0 \quad \sum \mu_i = 1 \]

EXPECTATION PARAMETER

\[ p(x | \mu) = \prod_{i=1}^{k} \mu_i^{x_i} \]

\[ -\log p(x | \mu) = -\sum_{i} x_i \ln \mu_i \]

\[ = \sum_{i} x_i \ln \frac{x_i}{\mu_i} \quad \text{REL. ENTROPY} \]

\[ D = (x^1, x^2, \ldots, x^t) \]

\[ k\text{-DIMENSIONAL UNIT VECTS} \]

\[ \arg \min_{\tilde{\mu}} -\log p(D | \tilde{\mu}) \]

\[ = \arg \min_{\tilde{\mu}} \sum_{t} -\log p(x^t | \tilde{\mu}) \]

\[ = \arg \min_{\tilde{\mu}} \sum_{i} \sum_{t} -x_i^{t} \ln \mu_i \]

\[ = \arg \min_{\tilde{\mu}} \sum_{i} \sum_{t} \frac{x_i^{t}}{t} \ln \frac{x_i^{t}}{\mu_i} \]

\[ = \arg \min_{\tilde{\mu}} \sum_{i} \sum_{t} \frac{x_i^{t}}{t} \ln \frac{x_i^{t}}{\mu_i} \]

RELATIVE ENTROPY
PROB UCES

\[ \text{RE}(\bar{a}, \bar{b}) \geq 0 \]
\[ \text{RE}(\bar{a}, \bar{b}) = 0 \quad \text{iff} \quad \bar{a} = \bar{b} \]

\[ \Rightarrow \quad \hat{m}_{LY} = \frac{\sum x_i^t}{T} \quad \text{EMPIRICAL COUNTS} \]

\[ \hat{m}_{LY} = \frac{1 + \sum x_i^t}{T + m} \quad \text{ONE ADDITIONAL EXAMPLE PER OUTCOME} \]
Example App:

Spam Filtering with "Naive Bayes"

Message d represented as bit vector

\((x_1, x_2, \ldots, x_m)\)

\(x_i = \begin{cases} 1 & \text{token } t_i \text{ in document} \\ 0 & \text{not in} \end{cases} \)

Each token \(t_i\) corresponds to a feature

Two categories

\(C_S\) Spam
\(C_H\) Ham

\[ P(C_S | \bar{X}) = \frac{P(C_S) P(\bar{X} | C_S)}{P(C_S) P(\bar{X} | C_S) + P(C_H) P(\bar{X} | C_H)} \]

Naive Bayes assumption

\[ P(X | C) = \prod P(t_i = x_i | C) \]

\(\uparrow\) Assumed Independence
\( P(t_i | c) \) MODELED AS BERNULLI

**NAIVE: TOKEN CO-OCCURRENCES**

IN CATEGORY NOT INDEPENDANT

\[
P(c_s | x) = \frac{\prod_{i=1}^{m} P(t_i | c_s) x_i (1 - P(t_i | c_s))^{1 - x_i}}{\sum_{c \in \{c_1, c_2, c_3\}} \prod_{i=1}^{m} P(t_i | c) x_i (1 - P(t_i | c))^{1 - x_i}}
\]

\( P(t_i | c) \) ESTIMATED USING LAPLACIAN PRIOR

\[
P(t_i | c) = \frac{1 + M_{t,i,c}}{2 + M_c}
\]

- \( M_{t,i,c} \) # OF TRAINING MESSAGES OF CATEGORY \( c \) THAT CONTAINS TOKEN \( t \)
- \( M_c \) # OF TRAINING MESSAGES OF CATEGORY \( c \)

DOCUMENT \( x \) CLASSIFIED AS SPAM IF

\[
P(c_s | x) > \text{THRESHOLD}
\]
ALTERNATE (MULTINOMIAL NAIVE BAYES)

MESSAGE I = BAG OF TOKENS
COUNTS MATTER

A REPRESENTED AS VECTOR

\[ \mathbf{x} = (x_1, \ldots, x_m) \text{, where } x_i \text{ is # of occurrences of } t_i \]

CATEGORY \( c \) AS DOING \( |d| = \sum_{i=1}^{m} x_i \) INDEPENDENT
DRAWS FROM MULTINOMIAL OVER \( (t_1, \ldots, t_m) \)

\[
P(\mathbf{x}|c) = \prod_{i=1}^{m} p(t_i|c)^{x_i}
\]

\[
P(c|\mathbf{x}) = \frac{P(c) \prod_{i=1}^{m} p(t_i|c)^{x_i}}{\sum_{c \in \{c_0, c_1, c_2\}} P(c) \prod_{i=1}^{m} p(t_i|c)^{x_i}}
\]

\[
p(t_i|c) = \frac{1 + N_{t_i,c}}{m + N_c}
\]

\( N_{t_i,c} \) # OF OCCURRENCES OF TOKEN \( t_i \) IN TRAINING MESSAGES OF CATEGORY \( c \)

\( N_c \) --

\( \mathbf{x} \) CLASSIFIED AS SPAM IF \( P(c_0|\mathbf{x}) > \text{THRESHOLD} \)