Notation:

\[ \mathcal{D} : \text{data set} \subseteq \{(x_i, y_i)\}_{i=1}^n \]

\[ \mathcal{D}_i : \text{data sample} \subseteq \{(x_i, y_i)\} \]

\[ \mathcal{X} \text{ : feature vector} \]

\( \theta \) : parameters of Bernoulli hypothesis for MLE in binary \( \theta \in \{0, 1\} \)

\( \mathcal{X}, \mathcal{Y} \) : design matrix, features of examples \( \mathcal{Y} = y \)

**Naive Bayes**

**Advantages**
- Simple
- Fast training
- Scalable
- Small per model
- Not much data needed

**Disadvantages**
- Non-linear dependence assumption
- Limited modeling power
- Only certain data types

Today's goal: \( x \in \mathbb{R}^d \) and which feature instead of binary

**Gaussian Distribution**

\[ x \sim \mathcal{N}(\mu, \Sigma) \]

\[ p(x | \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right) \]

**Motivation**
- Central Limit Theorem (converge in mean)
- Analytic properties (mean, variance, moments)
- Simple in sample
- Data in Gaussian

**MLE**

\[ \hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \]

\[ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})(x_i - \hat{\mu})^T \]

\[ p(x, y; \mathbf{\theta}, \phi) = \frac{1}{Z(\phi)} \exp\{\mathbf{\theta}^T \mathbf{G}(\mathbf{x}) + \phi(\mathbf{x})\} \]

where \( Z(\phi) \) is the normalization constant.

**Projected Idea**

- Explore ways to relax NB assumption
- (TAN Bayes, Bayesian network classifiers)
- More sophisticated
Gaussian Discriminant Analysis

\[ p(x_2 \mid y_2) = N(x_2 \mid \mu_2, \Sigma_2) \]

where \( \mu_2 \) and \( \Sigma_2 \) are the mean and covariance of the distribution of \( x_2 \) given \( y_2 \).

\[ p(x_1 \mid y_1) = N(x_1 \mid \mu_1, \Sigma_1) \]

This is the same for \( x_1 \) as well.

The discriminant function is given by

\[ 
D(x) = \frac{1}{2} [x - \mu_1] \Sigma_1^{-1} (x - \mu_2) + \frac{1}{2} \log \left| \frac{\Sigma_1}{\Sigma_2} \right| + \log p(y_1) - \log p(y_2) 
\]

In order to classify a new observation, we need to know the means and covariances of the two classes.

For linear discriminant analysis, we assume that each class follows a normal distribution with the same covariance matrix, \( \Sigma_1 = \Sigma_2 = \Sigma \).

Let \( X \) be the data matrix of \( N \) observations and \( n \) features, and let \( y \) be the class labels.

A linear discriminant analysis model is of the form

\[ Y = W^T X + b \]

where \( W \) is the weight vector and \( b \) is the bias term.

The goal is to find the weight vector \( W \) that maximizes the ratio of the between-class variance to the within-class variance.

\[ W = \arg \max_{W} \frac{\text{Tr}(W^T \Sigma W)}{\text{Tr}(W^T \Sigma W)} \]

subject to the condition that \( W^T \Sigma W = 1 \).

This is a constrained optimization problem, and it can be solved using the method of Lagrange multipliers.

The solution for \( W \) can be found using the following equation:

\[ W = \Sigma^{-1} \mu_Y - \Sigma^{-1} \mu_X \]

where \( \mu_Y \) and \( \mu_X \) are the mean vectors for the two classes.

The discriminant function can then be used to classify new observations.