

CS 3/5780

Logistic Regression

Spring 2026

Naive Bayes

- Idea: Model joint distribution $P(Y, X)$
- Naive Bayes Assumption: $P(X = \mathbf{x} | Y = y) = \prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha] | Y = y)$
- Predict using: $P(Y = y | X = \mathbf{x}) = \frac{\prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha] | Y = y) P(Y = y)}{\sum_{y' \in \mathcal{Y}} \prod_{\alpha=1}^d P(X[\alpha] = \mathbf{x}[\alpha] | Y = y') P(Y = y')}$

Conditional Distribution of Label

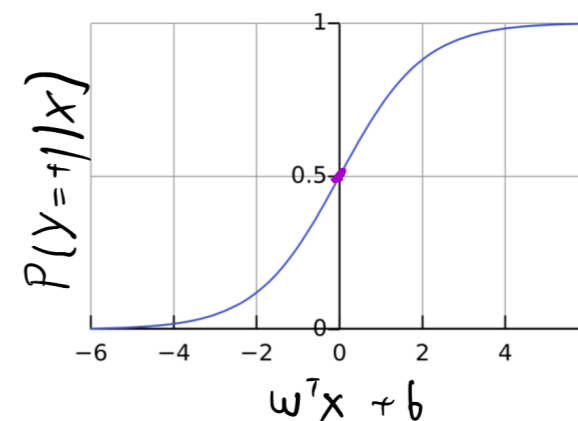
- Multinomial Naive Bayes
- Gaussian Naive Bayes

$$P(X = \mathbf{x} | Y = y) = \frac{m!}{\prod_{\alpha=1}^d x[\alpha]!} \prod_{\alpha=1}^d (\theta_{y[\alpha]})^{x[\alpha]} \propto \prod_{\alpha=1}^d (\theta_{y[\alpha]})^{x[\alpha]} \quad P(X[\alpha] = \mathbf{x}[\alpha] | Y = y) = \frac{1}{\sqrt{2\pi\sigma_{\alpha,y}^2}} \exp\left(-\frac{1}{2\sigma_{\alpha,y}^2} (x[\alpha] - \mu_{\alpha,y})^2\right)$$

$$P(Y = 1 | X = \mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^\top \mathbf{x} + b))} = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

Logistic Regression

$$P(Y = 1 | X = \mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^\top \mathbf{x} + b))} = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

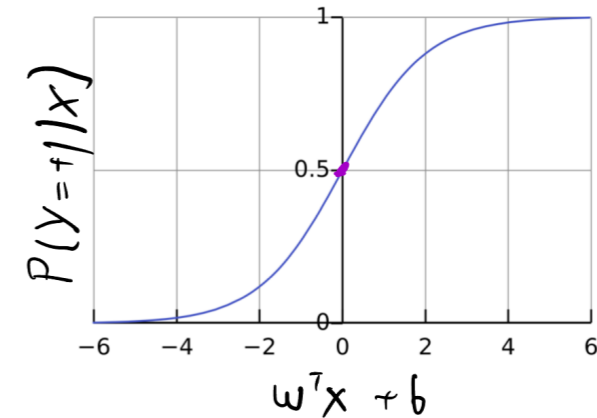


Generative Vs Discriminative Model

- Generative model we try to model the joint $P(X,Y)$ (Eg. Naive Bayes)
- But to finally classify we only need to evaluate $P(Y|X)$
- Discriminative Approach: Directly only model $P(Y|X)$
- Dont assume anything about $P(X)$ and avoid modeling it all together

Logistic Regression

$$P(Y = 1|X = \mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^\top \mathbf{x} + b))} = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$



MLE: Logistic Regression

$$\begin{aligned} \hat{\mathbf{w}}_{\text{MLE}} &= \arg \max_{\mathbf{w}} P(D|\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} P((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)|\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P((\mathbf{x}_i, y_i)|\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P(y_i|\mathbf{x}_i, \mathbf{w}) \times P(\mathbf{x}_i|\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P(y_i|\mathbf{x}_i, \mathbf{w}) \times P(\mathbf{x}_i) \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P(y_i|\mathbf{x}_i, \mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \sum_{i=1}^n \log(P(y_i|\mathbf{x}_i, \mathbf{w})) \\ &= \arg \max_{\mathbf{w}} \sum_{i=1}^n \log\left(\frac{1}{1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)}\right) \\ &= \arg \min_{\mathbf{w}} \sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) \end{aligned}$$

MAP: Logistic Regression

$$\begin{aligned} \hat{\mathbf{w}}_{\text{MAP}} &= \arg \max_{\mathbf{w}} P(\mathbf{w}|D) \\ &= \arg \max_{\mathbf{w}} P(D|\mathbf{w}) \times P(\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} P((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)|\mathbf{w}) \times P(\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \prod_{i=1}^n P((\mathbf{x}_i, y_i)|\mathbf{w}) \times P(\mathbf{w}) \\ &= \arg \max_{\mathbf{w}} \sum_{i=1}^n \log(P(y_i|\mathbf{x}_i, \mathbf{w})) + \log(P(\mathbf{w})) \\ &= \arg \min_{\mathbf{w}} \sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) - \log(P(\mathbf{w})) \end{aligned}$$

Summary

$$\ell(\mathbf{w}^\top \mathbf{x}, y) = \log(1 + \exp(-y\mathbf{w}^\top \mathbf{x}))$$

- How do we perform minimization:

$$\hat{\mathbf{w}}_{\text{MLE}} = \arg \min_{\mathbf{w}} \sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i))$$

- In next lecture we will look into iterative optimization methods like gradient descent, Newton's method and Stochastic Gradient Descent for such optimizations
- Logistic model easily extends to multi-class case as

$$P(Y = k | X = \mathbf{x}) = \frac{\exp(-\mathbf{w}_k^\top \mathbf{x})}{\sum_{j=1}^K \exp(-\mathbf{w}_j^\top \mathbf{x})}$$

for K parameters $\mathbf{w}_1, \dots, \mathbf{w}_K$ each in d dimensions.