The Naïve Bayes Classifier; Logistic Regression

Outline

- Review of the Bayes classifier and plug-in methods
- Naïve Bayes
- Logistic regression

Classification: Probabilistic Setting

- Feature vector $\boldsymbol{X} \in \mathbb{R}^d$
- Label $Y \in \{1, \ldots, K\}$
- Assume $(\boldsymbol{X}, Y)$ are jointly distributed ($d+1$ dimensional)
- A *classifier* is a function $f : \mathbb{R}^d \to \{1, \ldots, K\}$
- What is the best possible classifier?

Bayes Classifier

- The best classifier depends on the performance measure. The most common performance measure is the probability of error, or *risk*, defined by:

  $$R(f) := \Pr(f(\boldsymbol{X}) \neq Y)$$

  i.e., the probability of the event

  $$\{(\boldsymbol{x}, y) \in \mathbb{R}^d \times \{1, \ldots, K\} \mid f(\boldsymbol{x}) \neq y\}$$

- The *Bayes risk* is the smallest risk of any classifier, and is denoted R^*.

- If $R(f) = R^*$, f is called a *Bayes classifier*.

Bayes Classifier

- Notation:
  - $\pi_k := \Pr(Y = k)$ class prior
  - $g_k(\boldsymbol{x}) :=$ pdf/pmf of $\boldsymbol{X}$ given $Y = k$ class-conditional distribs.
  - $\eta_k(\boldsymbol{x}) := \Pr(Y = k | \boldsymbol{X} = \boldsymbol{x})$ class posterior
  - $g(\boldsymbol{x}) :=$ pdf/pmf of $\boldsymbol{X}$
- **Theorem:** The classifier

$$f^*(\boldsymbol{x}) = \underset{k=1,\ldots,K}{\arg\max} \quad \eta_k(\boldsymbol{x})$$
$$= \underset{k=1,\ldots,K}{\arg\max} \quad \pi_k g_k(\boldsymbol{x})$$

is a Bayes classifier.

Review: Plug-In Classifiers

- In machine learning, the joint distribution of $(\boldsymbol{X}, Y)$ (as captured by π_k and g_k, or g and η_k) is not known, so we can't know the Bayes' classifier.

- However, the formula for the Bayes' classifier is still useful. We can estimate the quantities in the formula from training data, and plug those estimates in to the formula to get a classifier.

- Linear discriminant analysis and Naïve Bayes have the form

$$\widehat{f}(\boldsymbol{x}) := \arg\max_k \widehat{\pi}_k \widehat{g}_k(\boldsymbol{x})$$

- Logistic regression has the form

$$\widehat{f}(\boldsymbol{x}) := \arg\max_k \widehat{\eta}_k(\boldsymbol{x})$$

Naïve Bayes Assumption

- Training data:

$$(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \stackrel{iid}{\sim} P_{\boldsymbol{X}Y}.$$

- Notation:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_d \end{bmatrix}$$

- *Naïve Bayes assumption*: Given Y, the features $X_1, \ldots, X_d$ are

independent

Naïve Bayes Classifier

- Recall $g_k(\boldsymbol{x})$ is the pmf/pdf of $\boldsymbol{X}|Y = k$. By the Naïve Bayes assumption,

$$\boxed{g_k(\boldsymbol{x}) = \prod_{j=1}^{d} g_{kj}(x_j)}$$

where $g_{kj}(x_j)$ is the marginal pmf/pdf of $X_j \mid Y = k$.

- Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be training data, and let

$\widehat{\pi}_k =$ proportion of class k in the training data

$\widehat{g}_{kj} =$ estimate of g_{kj}

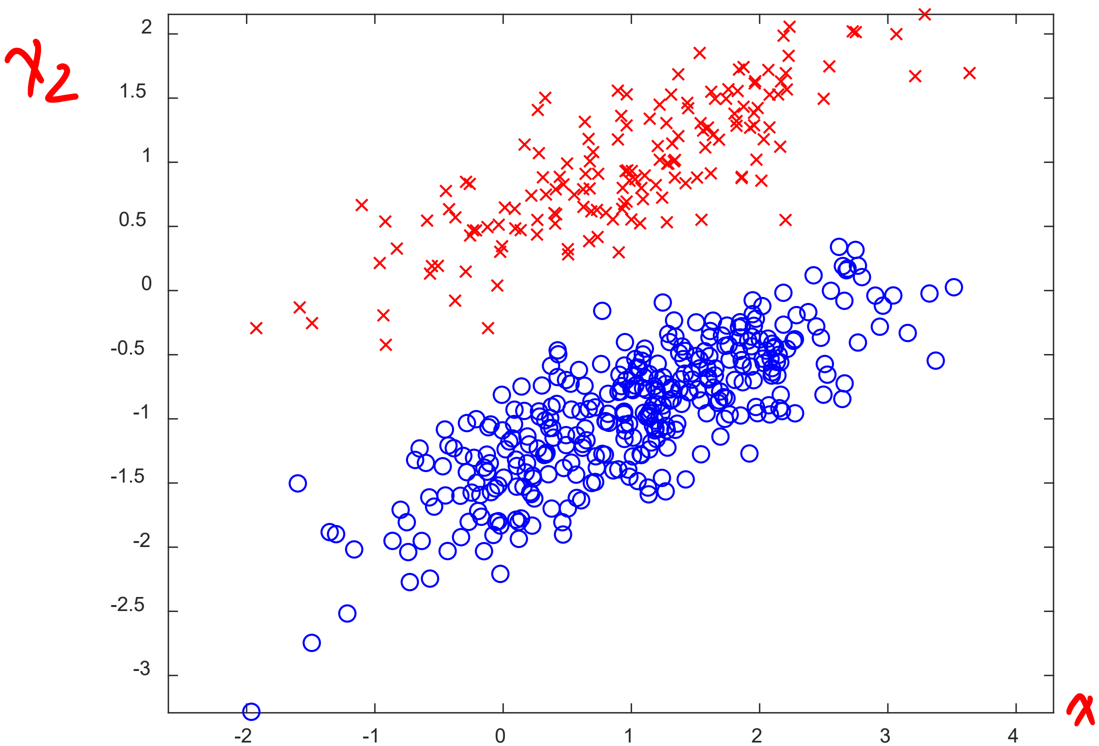
- Then the Naïve Bayes classifier is

$$\widehat{f}(\boldsymbol{x}) = \text{argmax}_k \; \hat{\pi}_k \hat{g}_k(x) = \underset{k}{\arg\max} \; \hat{\pi}_k \prod_{j=1}^{d} \hat{g}_{kj}(x_j)$$

Example: Gaussian Data

$g_1(x) = g_{11}(x_1)\, g_{12}(x_2)$

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$



$g_{11}(x_1) = \phi(x_1;\, \mu_{11}, \sigma_{11}^2)$

$\hat{g}_{11}(x_1) = \phi(x_1;\, \hat{\mu}_{11}, \hat{\sigma}_{11}^2)$,

$\hat{\mu}_{11} = \frac{1}{n_1} \sum_{i:y_i=1} x_{i1}$

$\hat{\sigma}_{11}^2 = \frac{1}{n_1} \sum_{i:y_i=1} (x_{i1} - \hat{\mu}_{11})^2$

Document Classification

- Suppose we wish to classify documents into categories like "business," "politics," "sports," etc.
- A simple yet popular feature representation is the *bag-of-words* representation.
- A document is represented as a vector

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

where d is the number of words in the vocabulary, and

$$x_j = \text{ number of times word } j \text{ occurs in document}$$

- $\widehat{g}_{kj}(\ell) =$ proportion of times that x_j occurs exactly ℓ times among training documents from class k.

Poll

True or False: The Naïve Bayes assumption is a reasonable assumption for document classification with a bag of words representation

(A) True

(B) False ✓

Logistic Regression

- Focus on binary case, $Y \in \{0, 1\}$
- Logistic regression produces an estimate $\widehat{\eta}(\boldsymbol{x})$ of

$$\eta(\boldsymbol{x}) := \Pr(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$$

- As such, it is a plug-in classifier,

$$\widehat{f}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \hat{\eta}(x) > \frac{1}{2} \\ 0 & \text{if } \hat{\eta}(x) \leq \frac{1}{2} \end{cases}$$

- It combines two ideas:
  - Logistic probability estimation
  - Linear classification

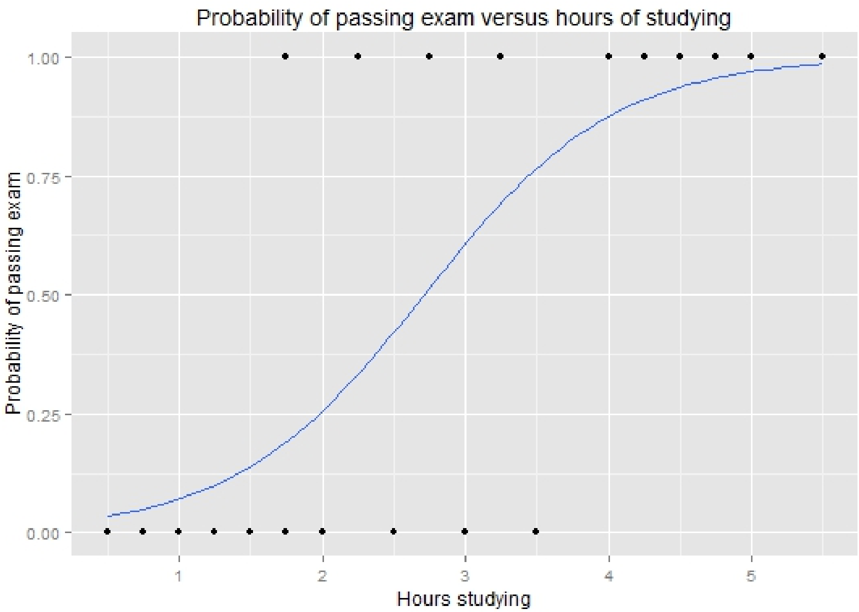
Logistic Probability Estimation

- Example taken from Wikipedia

  A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

$$\Pr(\text{Pass} \mid \text{Hours}) = \frac{1}{1 + \exp(-(1.5 \times \text{Hours} - 4))}$$



- Prediction
- Estimation
- Multi-dimensional extension? E.g., predict label based on multiple attributes

Logistic Regression Model

- Feed linear combination of features into logistic probability model

$$\eta(\boldsymbol{x};\boldsymbol{\theta}) := \frac{1}{1+\exp(-[w^T x + b])}, \quad w \in \mathbb{R}^d, \; b \in \mathbb{R}$$

where

$$\boldsymbol{\theta} = \begin{bmatrix} b \\ w_1 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$\hat{\eta}(x) = \eta(x;\hat{\theta})$$

- Plug-in method is a linear classifier

$$\hat{\theta} = \begin{bmatrix} \hat{b} \\ \hat{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

predict class 1 $\iff \hat{\eta}(x) > \frac{1}{2}$

$\iff \eta(x;\hat{\theta}) > \frac{1}{2}$

$\iff \hat{w}^T x + \hat{b} > 0$

Parameter Estimation

- Given training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, how should we set

$$\boldsymbol{\theta} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}?$$

We need a criterion that quantifies how well

$$\eta(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{1 + \exp(-[\boldsymbol{w}^T \boldsymbol{x} + b])}$$

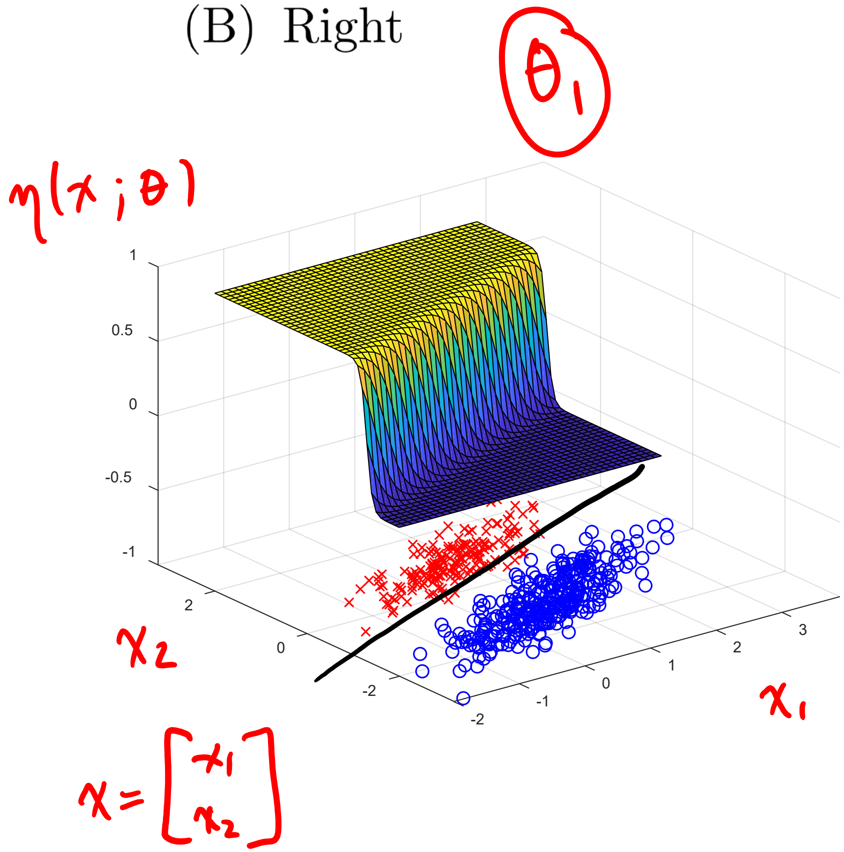
explains the training data for each $\boldsymbol{\theta}$.

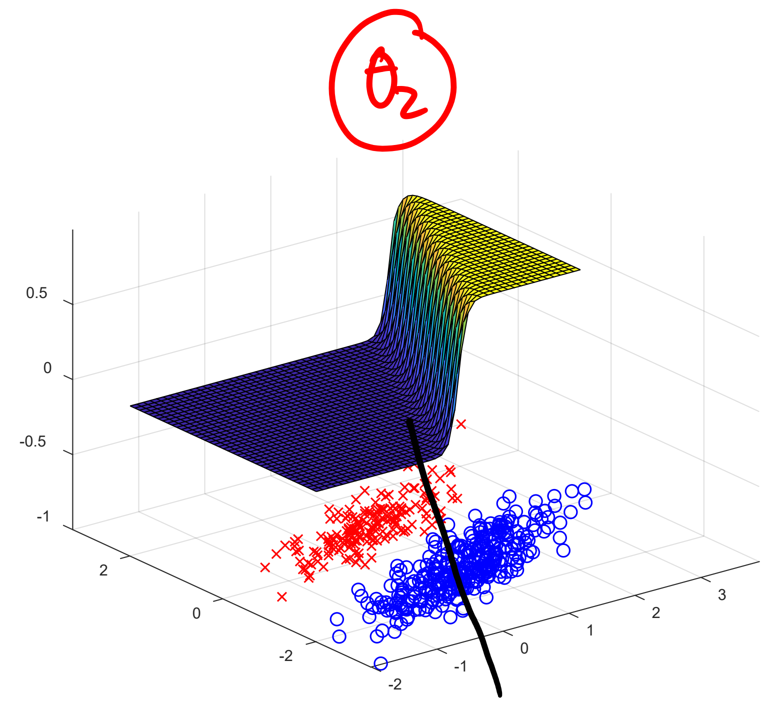
Poll

Which choice of $\boldsymbol{\theta}$ provides the better fit to the data shown?

(A) Left

(B) Right





Parameter Estimation

- Let $p(y \mid \boldsymbol{x}; \boldsymbol{\theta})$ denote the conditional pmf of y given x.
- Observe

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \begin{cases} \eta(x;\theta) & y=1 \\ 1-\eta(x;\theta) & y=0 \end{cases}$$

$$= \eta(x;\theta)^{y} \left(1-\eta(x;\theta)\right)^{1-y}$$

- The *likelihood* of $\boldsymbol{\theta}$ is defined to be

$$L(\boldsymbol{\theta}) := \prod_{i=1}^{n} p(y_i \mid x_i;\theta)$$

$$= \prod_{i=1}^{n} \eta(x_i;\theta)^{y_i} \left(1-\eta(x_i;\theta)\right)^{1-y_i}$$

Log Likelihood

- Notation: $w^T x_i + b \Leftrightarrow \theta^T \tilde{x}_i$

$$\tilde{\boldsymbol{x}}_i = [1 \; x_{i1} \; \cdots \; x_{id}]^T$$
$$\boldsymbol{\theta} = [b \; w_1 \; \cdots \; w_d]^T$$

- Then the likelihood can be expressed

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left(\frac{1}{1+\exp(-\theta^T \tilde{x}_i)} \right)^{y_i} \left(\frac{e^{-\theta^T \tilde{x}_i}}{1+e^{-\theta^T \tilde{x}_i}} \right)^{1-y_i}$$

- The *log-likelihood* of $\boldsymbol{\theta}$ is

$$\ell(\boldsymbol{\theta}) := \log L(\boldsymbol{\theta})$$
$$= \sum_{i=1}^{n} \left[y_i \log \left(\frac{1}{1+e^{-\theta^T \tilde{x}_i}} \right) + (1-y_i) \log \left(\frac{e^{-\theta^T \tilde{x}_i}}{1+e^{-\theta^T \tilde{x}_i}} \right) \right]$$

Regularized Logistic Regression

- Unless $n \gg d$, it is preferable to minimize the modified objective function
$$J(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \lambda \|\boldsymbol{w}\|^2,$$
where $\lambda > 0$ is a fixed, used-specified constant called a *regularization parameter*.

- Why introduce the regularization term?

Iterative Optimization

- Unless $n \gg d$, it is preferable to minimize the modified objective function
$$J(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \lambda \|\boldsymbol{w}\|^2,$$
where $\lambda > 0$ is a fixed, used-specified constant called a *regularization parameter*.

- $\nabla J(\boldsymbol{\theta}) = \mathbf{0}$ cannot be solved analytically

- However, $J(\boldsymbol{\theta})$ is convex

- Several options for iterative algorithms
  - Gradient descent
  - Stochastic gradient descent
  - Newton's method
  - Majorize-minimize
  - ...

Logistic Regression Recap

- More than a classifier – it predicts the probability of each class
- LR assumption is less restrictive than LDA assumption
- Widely used in health sciences and other application domains
- Example: predict whether a patient will develop a disease (e.g., diabetes, coronary disease) based on various attributes (age, sex, weight, BMI, etc.)
- Multiclass extension: will discuss later
- Nonlinear extension:

$$\eta(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{1}{1 + \exp(-f_{\boldsymbol{\theta}}(\boldsymbol{x}))}$$

  where f is nonlinear. We will see examples (kernel methods, neural networks)