Generative Models for Classification: Linear Discriminant Analysis (LDA) and Naive Bayes

Sujay Sanghavi

Recall: Probabilistic View of Logistic Regression

Features and targets are random variables: X and Y respectively

Parameter β determines how X is related to Y

$$\mathbf{P}_\beta(Y = 1|X = x) \;=\; \frac{1}{1 + \exp(-\beta^\top x)}$$

Given data (x, y), find $\widehat{\beta}$ by maximizing

$$\widehat{\beta} \;=\; \arg\max_\beta \mathbf{P}_\beta(Y = y \,|\, X = x)$$

Thus we find the β that best models the conditional distribution of the label given the features.

But this is not the only way one can think about this problem ...

Two Paths to Same Goal

The joint distribution of the features and the data is $\mathbf{P}(X = x, Y = y)$

This can be split two ways. Each way gives a different approach.

(A) $\mathbf{P}(X = x, Y = y) = \mathbf{P}(Y = y|X = x)\mathbf{P}(X = x)$

This corresponds to discriminative methods, like logistic regression

(B) $\mathbf{P}(X = x, Y = y) = \mathbf{P}(X = x|Y = y)\mathbf{P}(Y = y)$

This corresponds to generative methods

Discriminative vs Generative Models

Discriminative classifier: where β models $\mathbf{P}_\beta(Y = y \,|\, X = x)$, but does not care about $\mathbf{P}(X = x)$, i.e. how the features were generated

$$\mathbf{P}_\beta(Y = y\,, X = x) \;=\; \mathbf{P}_\beta(Y = y \,|\, X = x)\, \mathbf{P}(X = x)$$

(e.g. Logistic Regression)

Generative classifier: where β models the feature distribution for each class, i.e. $\mathbf{P}_\beta(X = x \,|\, Y = y)$ and then combines it with $\mathbf{P}(Y = y)$

$$\mathbf{P}_\beta(Y = y\,, X = x) \;=\; \mathbf{P}_\beta(X = x \,|\, Y = y)\, \mathbf{P}(Y = y)$$

Here $\mathbf{P}(Y = y)$ is often just fraction of training samples with that label ...

Discriminative models learn $\mathbf{P}_\beta(Y = y \,|\, X = x)$ while generative models learn $\mathbf{P}_\beta(X = x \,|\, Y = y)$

Linear Discriminant Analysis (LDA)

A simple **generative** classifier. The class distribution $\mathbf{P}_\beta(X = x \mid Y = y)$ is assumed to be a **Gaussian.**

If binary classification, i.e. labels y can be 0 or 1, then

$$\beta \triangleq (\mu_0, \mu_1, \pi_0, \pi_1, \Sigma)$$

so that

$$\begin{aligned} \mathbf{P}_\beta(X = x \mid Y = 0) &\quad \text{is the Gaussian} \quad \mathcal{N}(\mu_0, \Sigma) \\ \mathbf{P}_\beta(X = x \mid Y = 1) &\quad \text{is the Gaussian} \quad \mathcal{N}(\mu_1, \Sigma) \\ \mathbf{P}_\beta(Y = 0) &\quad \text{is given by} \quad \pi_0 \\ \mathbf{P}_\beta(Y = 1) &\quad \text{is given by} \quad \pi_1 \end{aligned}$$

$\widehat{\beta}$, i.e. $\widehat{\mu_0}, \widehat{\mu_1}, \widehat{\pi_0}, \widehat{\pi_1}, \widehat{\Sigma}$ are estimated from data

Linear Discriminant Analysis (LDA)

Q1: How do we find $(\widehat{\mu_0}, \widehat{\mu_1}, \widehat{\pi_0}, \widehat{\pi_1}, \widehat{\Sigma})$ from data

A1: In the natural way:

$$\begin{aligned} \widehat{\mu_0} &= \text{center of all samples with label 0} \\ \widehat{\mu_1} &= \text{center of all samples with label 1} \\ \widehat{\pi_0} &= \text{fraction of samples with label 0} \\ \widehat{\pi_1} &= \text{fraction of samples with label 1} \end{aligned}$$

and $\widehat{\Sigma}$ is the covariance estimate of (each of the) Gaussians (they are both assumed to have same covariance)

Linear Discriminant Analysis (LDA)

Q2: Given a model $\beta = (\mu_0, \mu_1, \pi_0, \pi_1, \Sigma)$, what is the optimal (not necessarily linear) classifier ?

A2: For any new point x, find which of the two Gaussians in β was the most likely to have generated that x, i.e.

$$\begin{aligned} \widehat{y}(x) &= 0 \text{ if } \mathbf{P}_\beta(X = x|Y = 0)\mathbf{P}(Y = 0) > \mathbf{P}_\beta(X = x|Y = 1)\mathbf{P}(Y = 1) \\ &= 1 \text{ else} \end{aligned}$$

Linear Discriminant Analysis (LDA)

Q2: Given a model $\beta = (\mu_0, \mu_1, \pi_0, \pi_1, \Sigma)$, what is the optimal (not necessarily linear) classifier ?

A2: For any new point x, find which of the two Gaussians in β was the most likely to have generated that x, i.e.

$$\begin{aligned} \widehat{y}(x) &= 0 \text{ if } \mathbf{P}_\beta(X = x|Y = 0)\mathbf{P}(Y = 0) > \mathbf{P}_\beta(X = x|Y = 1)\mathbf{P}(Y = 1) \\ &= 1 \text{ else} \end{aligned}$$

This can be reduced to: for $k = 0$ and $k = 1$ define

$$\delta_k(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log \pi_k$$

and then

$$\begin{aligned} \widehat{y}(x) &= 0 \quad \text{if } \delta_0(x) > \delta_1(x) \\ &= 1 \quad \text{else} \end{aligned}$$

$\delta_k(x)$ is the **"discriminant"** and it is a **linear** function of x. Hence, **LDA**.

Logistic Regression vs LDA

Both methods are **linear** classifiers: there is a vector v such that

$$\widehat{y}(x) \;=\; 1 \quad \Leftrightarrow \quad v^\top x > 0$$

Logistic regression: $v = \widehat{\beta}$

LDA: v can be found by simplifying the rule $\delta_0(x) > \delta_1(x)$

However, the v is learnt differently in the two cases.

Both Logistic Regression and LDA can be extended to more than two classes ... both still keep linear structure.

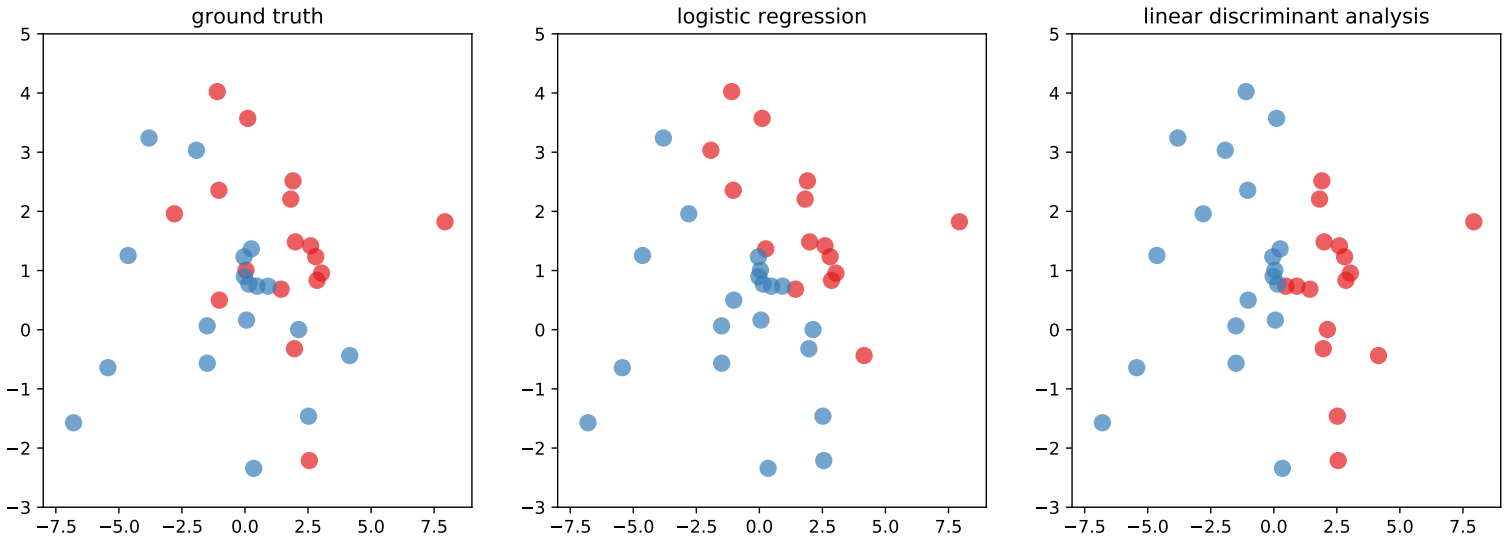
Logistic Regression vs LDA

Both are **linear classifiers**, but learnt differently from the data.

LDA: Assumes $\mathbf{P}(x|y)$ is a Gaussian. So, performance may suffer if truth is far from this assumption.

Logistic Regression: The sigmoid function becomes almost flat far from 0; results in instability if most data is very well-separated.

Example when LR is better than LDA

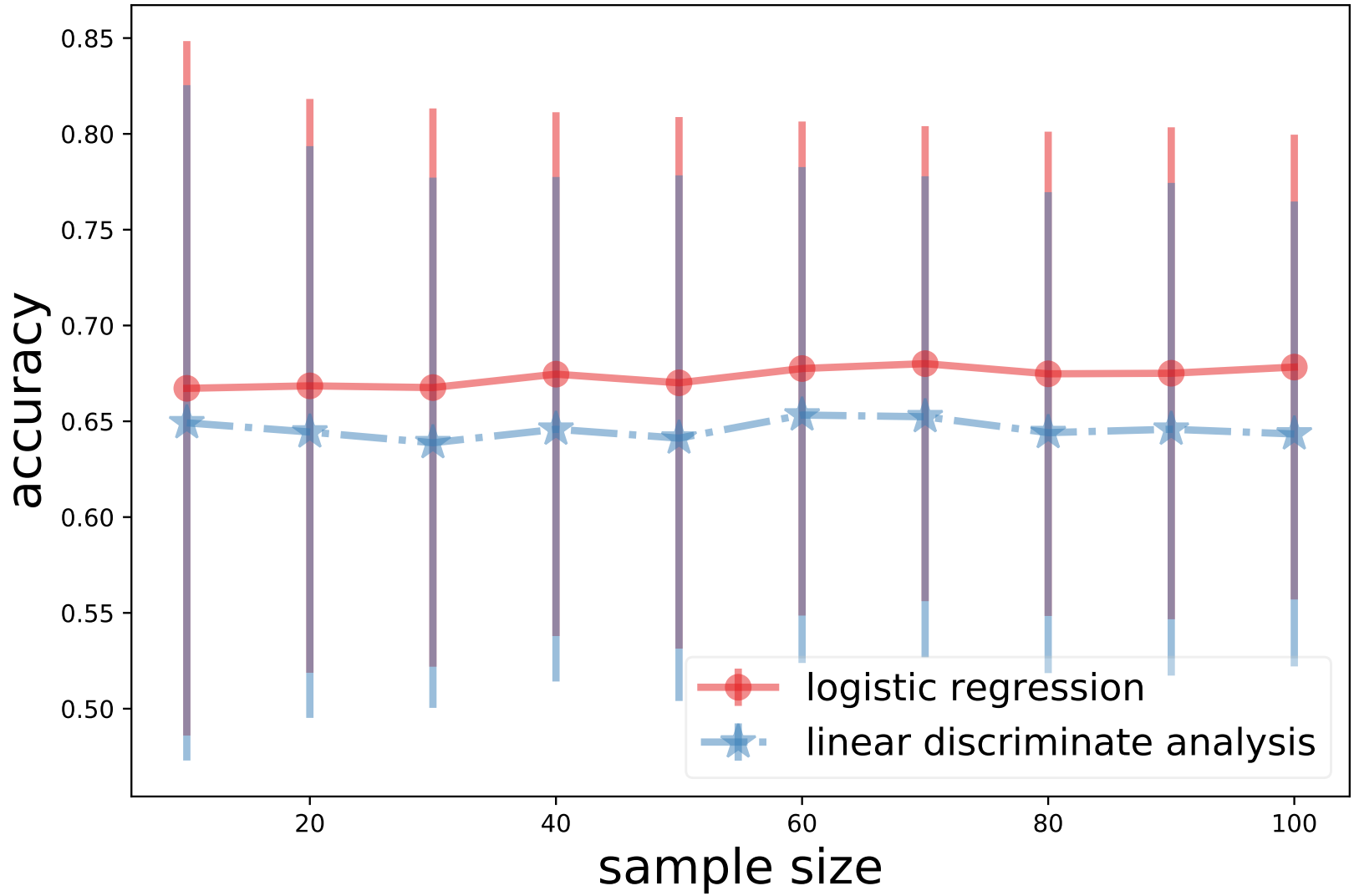


In this example:

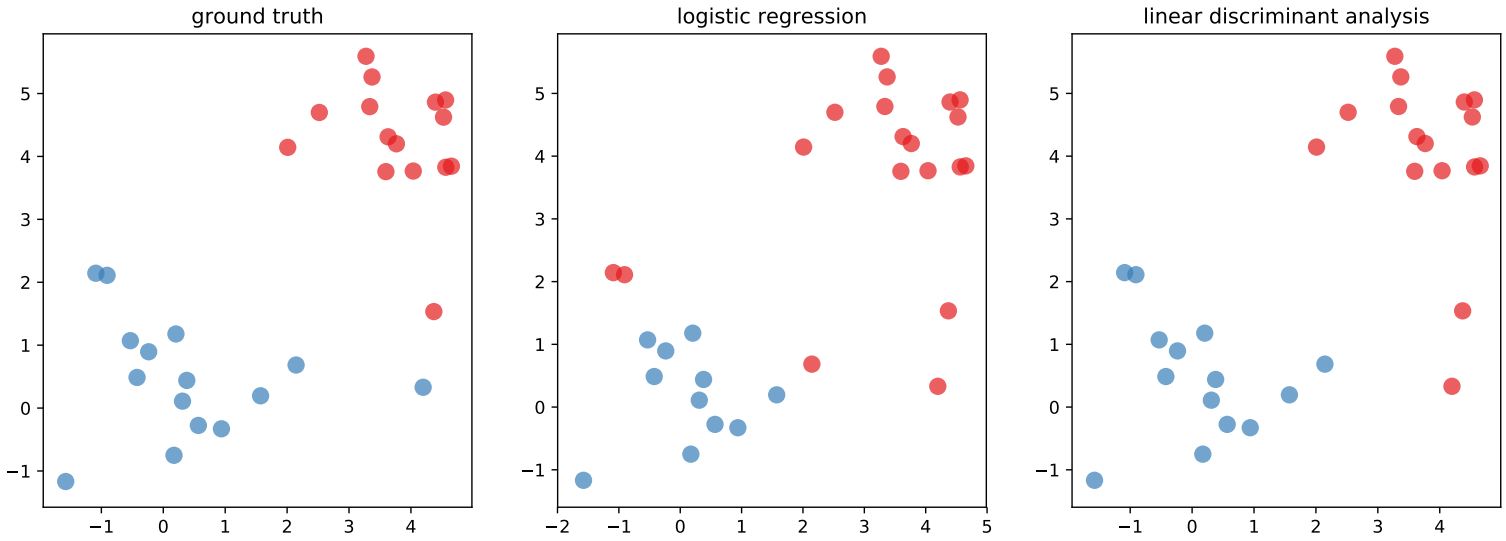
For $i \in C_0$, $x_{i,1}, x_{i,2}$ follow standard Cauchy distribution with mode 0;

For $i \in C_1$, $x_{i,1}, x_{i,2}$ follow standard Cauchy distribution with mode 2;

Example when LR is better than LDA



Examples when LDA is better than LR

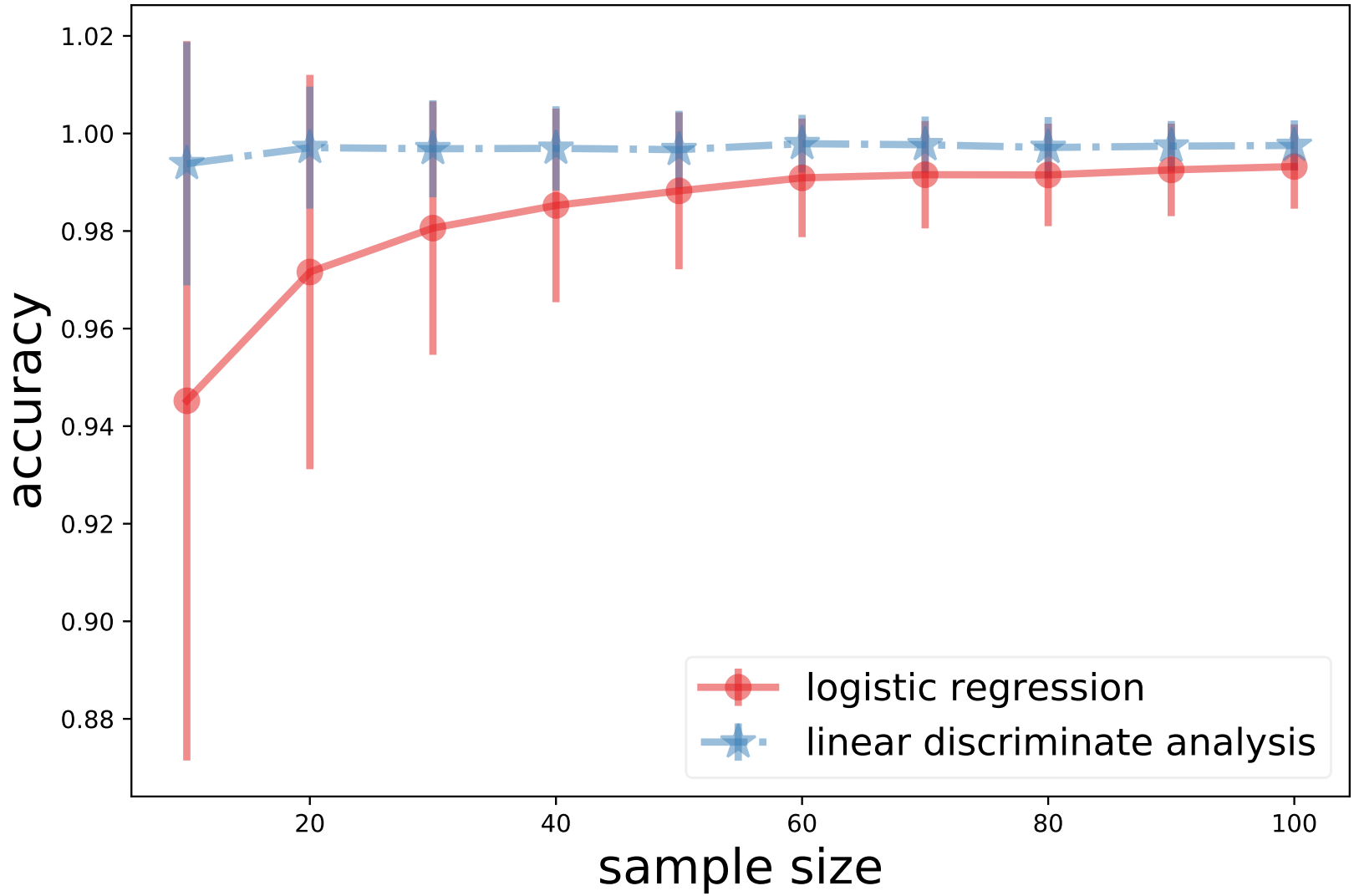


In this example:

For $i \in C_0$, $x_{i,1}, x_{i,2}$ follow standard Gaussian distribution with mean 0;

For $i \in C_1$, $x_{i,1}, x_{i,2}$ follow standard Gaussian distribution with mean 4;

Examples when LDA is better than LR



Naive Bayes

A **generative model for classification** that assumes that

$$\mathbf{P}\left((x_1, \ldots, x_d) \,|\, y\right) \;=\; \mathbf{P}(x_1|y)\,\mathbf{P}(x_2|y)\,\ldots\,\mathbf{P}(x_d|y)$$

i.e. **the features are conditionally independent given the label.**

Naive Bayes

A **generative model for classification** that assumes that

$$\mathbf{P}\left((x_1, \ldots, x_d) \,|\, y\right) \;=\; \mathbf{P}(x_1|y)\,\mathbf{P}(x_2|y)\,\ldots\,\mathbf{P}(x_d|y)$$

i.e. **the features are conditionally independent given the label.**

E.g. credit approval / denial

$$x = (\text{SeriousDlquin2yrs}, \text{Age}, \text{DebtRatio}, \text{MonthlyIncome}\ldots)$$

Naive Bayes assumes that the condiional distribution

$$\mathbf{P}\left((\text{SeriousDlquin2yrs}, \text{Age}, \text{DebtRatio}, \text{MonthlyIncome}\ldots)|\text{approved}\right)$$

is equal to the product

$$\mathbf{P}(\text{SeriousDlquin2yrs}|\text{approved}) \times \mathbf{P}(\text{Age}|\text{approved}) \times \mathbf{P}(\text{DebtRatio}|\text{approved}) \times$$

Naive Bayes

A **generative model for classification** that assumes that

$$\mathbf{P}\left((x_1, \ldots, x_d) \,|\, y \right) \;=\; \mathbf{P}(x_1|y)\, \mathbf{P}(x_2|y) \,\ldots\, \mathbf{P}(x_d|y)$$

i.e. **the features are conditionally independent given the label.**

Naive Bayes

A **generative model for classification** that assumes that

$$\mathbf{P}\left((x_1, \ldots, x_d) \,|\, y\right) \;=\; \mathbf{P}(x_1|y)\,\mathbf{P}(x_2|y)\,\ldots\,\mathbf{P}(x_d|y)$$

i.e. **the features are conditionally independent given the label.**

In particular, pick a **model class** $P_{\beta_j}(x_j|y)$ for each feature j in $1, \ldots, d$

This can be different for different features. So, some features can be Gaussian, others binary, etc.

Naive Bayes

A **generative model for classification** that assumes that

$$\mathbf{P}\left((x_1, \ldots, x_d) \,|\, y\right) \;=\; \mathbf{P}(x_1|y)\,\mathbf{P}(x_2|y)\,\ldots\,\mathbf{P}(x_d|y)$$

i.e. **the features are conditionally independent given the label.**

In particular, pick a **model class** $P_{\beta_j}(x_j|y)$ for each feature j in $1, \ldots, d$

This can be different for different features. So, some features can be Gaussian, others binary, etc.

Then, overall likelihood is assumed to be

$$\mathbf{P}_{\beta}\left(x \,|\, y\right) \;=\; \mathbf{P}_{\beta_1}(x_1|y) \times \mathbf{P}_{\beta_2}(x_2|y) \times \ldots \times \mathbf{P}_{\beta_d}(x_d|y)$$

Naive Bayes

Naive Bayes Training: Given training data $(x^{(i)}, y^{(i)})$ for $i = 1, \ldots, n$

(1) Find a model class $P_{\beta_j}(x_j|y)$ for the conditional distribution of each feature x_j given the label y.

(2) For each feature, individually, find the best $\widehat{\beta}_j$ by fitting to data:

$$\widehat{\beta}_j \;=\; \arg\max_{\beta_j} \prod_{i=1}^{n} \mathbf{P}_{\beta_j}(x_j^{(i)}|y^{(i)})$$

Naive Bayes

Naive Bayes Training: Given training data $(x^{(i)}, y^{(i)})$ for $i = 1, \ldots, n$

(1) Find a model class $P_{\beta_j}(x_j|y)$ for the conditional distribution of each feature x_j given the label y.

(2) For each feature, individually, find the best $\widehat{\beta}_j$ by fitting to data:

$$\widehat{\beta}_j \;=\; \arg\max_{\beta_j} \prod_{i=1}^{n} \mathbf{P}_{\beta_j}(x_j^{(i)}|y^{(i)})$$

Naive Bayes Inference: Given model $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_d)$ and new point x

(1) Evaluate the likelihood of each label y: $\mathbf{P}_{\widehat{\beta}}(x|y) \;=\; \prod_{j=1}^{d} \mathbf{P}_{\widehat{\beta}_j}(x_j|y)$

(2) Find the best label $\widehat{y}(x) \;=\; \arg\max_y \mathbf{P}_{\widehat{\beta}}(x|y)\mathbf{P}(y)$

Naive Bayes

A **generative model for classification** that assumes that

$$\mathbf{P}\left((x_1, \ldots, x_d) \,|\, y\right) \;=\; \mathbf{P}(x_1|y)\,\mathbf{P}(x_2|y)\,\ldots\,\mathbf{P}(x_d|y)$$

i.e. **the features are conditionally independent given the label.**

- Assumption avoids curse of dimensionality
- Each of the $\mathbf{P}(x_j|y)$ found from data – typically by fitting a *parametric model* (e.g. a 1-dimensional gaussian)
- Allows for a mix of features numerical and categorical features, since each feature is independent.

As in any generative method, once $\mathbf{P}(x|y)$ is estimated, classification by

$$\widehat{y}(x) \;=\; \arg\max_y \mathbf{P}(y)\,\mathbf{P}(x|y)$$

(Gaussian) Naive Bayes vs LDA

Gaussian Naive Bayes: for each feature x_i and label y estimate conditional mean $\widehat{\mu}_{i,y}$ and variance $\widehat{\sigma}^2_{i,y}$.

$$\widehat{y}(x) \;=\; \arg\max_y \; \mathbf{P}(y) \prod_i \frac{1}{\sqrt{2\pi \det(D)}} \exp\left(\frac{(x_i - \widehat{\mu}_{i,y})^2}{2\widehat{\sigma}^2_{i,y}}\right)$$

(Gaussian) Naive Bayes vs LDA

Gaussian Naive Bayes: for each feature x_i and label y estimate conditional mean $\widehat{\mu}_{i,y}$ and variance $\widehat{\sigma}^2_{i,y}$.

$$\widehat{y}(x) \;=\; \arg\max_y \; \mathbf{P}(y) \prod_i \frac{1}{\sqrt{2\pi \det(D)}} \exp\left(\frac{(x_i - \widehat{\mu}_{i,y})^2}{2\widehat{\sigma}^2_{i,y}}\right)$$

This can be rewritten as

$$\widehat{y}(x) \;=\; \arg\max_y \; \mathbf{P}(y) \frac{1}{\sqrt{2\pi\widehat{\sigma}^2_{i,y}}} \exp\left(-\frac{(x - \widehat{\mu}_y)^\top D^{-1}(x - \widehat{\mu}_y)}{2}\right)$$

(Gaussian) Naive Bayes vs LDA

LDA: assume classes share covariance Σ, and find per-class mean $\widehat{\mu}_y$ for each label y

$$\widehat{y}(x) \;=\; \arg\max_y \; \mathbf{P}(y) \; \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \; \exp\left(-\frac{(x-\widehat{\mu}_y)^\top \Sigma^{-1} (x-\widehat{\mu}_y)}{2} \right)$$

LDA allows for non-diagonal covariances, but forces them to be the same for all classes.
Gausian Naive Bayes only allows diagonal covariances, but these can be different for different classes.

Example of Naive Bayes

We want to classify whether a patient has **diabetes** based on:

- **Age** (continuous)
- **Exercise Level** (discrete: "Low", "Medium", "High")

The dataset is as follows:

Age (Years)	Exercise Level	Diabetes (Target)
25	Low	No
40	Medium	Yes
35	High	No
50	Low	Yes
45	Medium	Yes
60	High	No

This problem has two possible labels – "Yes" and "No" – which have to be predicted from 2 features.

Feature Distributions

The first step in Naive Bayes is to find the class-conditional distribution $\mathbf{P}(x_i|y)$ for every feature x_i given label y

Age is a continuous feature so **we will assume** it's class conditional distribution(s) are Gaussian.

$\mu_{\text{Age,Yes}} = (40 + 50 + 45)/3 = 45, \quad \sigma^2_{\text{Age,Yes}} = 16.67$

$\mu_{\text{Age,No}} = 40, \quad \sigma^2_{\text{Age,No}} = 216.67$

Exercise Level' is a discrete feature so we will just do direct counting:

$$P(\text{Exercise} = x|\text{Diabetes} = \text{Yes}) = \begin{cases} \frac{1}{3} & \text{if } x = \text{Low} \\ \frac{2}{3} & \text{if } x = \text{Medium} \\ 0 & \text{if } x = \text{High} \end{cases}$$

$$P(\text{Exercise} = x|\text{Diabetes} = \text{No}) = \begin{cases} \frac{1}{3} & \text{if } x = \text{Low} \\ 0 & \text{if } x = \text{Medium} \\ \frac{2}{3} & \text{if } x = \text{High} \end{cases}$$

Prediction for a New Patient

We want to predict whether a patient has diabetes given:

- Age $= 30$
- Exercise Level $=$ Low

Compute Gaussian Likelihood for Age
For **Diabetes = Yes**:

$$P(\text{Age} = 30 | \mu_{\text{Age,Yes}} = 45, \sigma^2_{\text{Age,Yes}} = 16.67) = \frac{1}{\sqrt{2\pi \cdot 16.67}} e^{-\frac{(30-45)^2}{2 \cdot 16.67}} \approx 0.0001$$

For **Diabetes = No**:

$$P(\text{Age} = 30 | \mu_{\text{Age,No}} = 40, \sigma^2_{\text{Age,No}} = 216.67) = \frac{1}{\sqrt{2\pi \cdot 216.67}} e^{-\frac{(30-40)^2}{2 \cdot 216.67}} \approx 0.022$$

Prediction for a New Patient

Compute Probabilities for Exercise category

For **Diabetes = Yes**:

$$P(\text{Exercise} = \text{Low}|\text{Diabetes} = \text{Yes}) = \frac{1}{3}$$

For **Diabetes = No**:

$$P(\text{Exercise} = \text{Low}|\text{Diabetes} = \text{No}) = \frac{1}{3}$$

Final Probabilities

For Diabetes = Yes:

$$P(\text{Diabetes=Yes}|\text{Age=30, Exercise=Low}) =$$

$$P(\text{Age=30}|\text{Diabates=Yes}) \cdot P(\text{Exercise=Low}|\text{Diabetes=Yes}) \cdot P(\text{Diabetes=Yes})$$

$$= 0.0001 \times \frac{1}{3} \times 0.5 = 1.67 \times 10^{-5}$$

For Diabetes = No:

$$P(\text{Diabetes=No}|\text{Age=30, Exercise=Low}) =$$

$$P(\text{Age=30}|\text{Diabetes=No}) \cdot P(\text{Exercise=Low}|\text{Diabetes=No}) \cdot P(\text{Diabetes=No})$$

$$= 0.022 \times \frac{1}{3} \times 0.5 = 0.00367$$

Final Prediction: No