Machine learning

Claire Boyer

September 16th, 2020

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

ISUP

Today

1. Mathematical framework
2. Discriminant analysis
   - The multivariate normal distribution
   - Bayes classifier for multivariate normal distributions
   - Rayleigh quotient
   - Fisher discriminant analysis
3. Logistic regression

Summary

1. Mathematical framework

2. Discriminant analysis
   - The multivariate normal distribution
   - Bayes classifier for multivariate normal distributions
   - Rayleigh quotient
   - Fisher discriminant analysis

3. Logistic regression

Mathematical framework for today

Supervised learning

Given observations, $d_n = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ we want to explain/predict outputs $y_i \in \mathcal{Y}$ from inputs $x_i \in \mathcal{X}$.

Goal

- Explain/Learn connections between inputs x_i and outputs y_i ;
- Predict the output y for a new input $x \in \mathcal{X}$

To do so, we have to find a machine or function $f : \mathcal{X} \to \mathcal{Y}$ such that

$$f(x_i) \simeq y_i, i = 1, \dots, n.$$

Jargon

- When the output y is continuous, $\rightsquigarrow$ regression problem
- When the output y is categorical, $\rightsquigarrow$ classification problem

Summary

1. Mathematical framework

2. Discriminant analysis
   - The multivariate normal distribution
   - Bayes classifier for multivariate normal distributions
   - Rayleigh quotient
   - Fisher discriminant analysis

3. Logistic regression

Recall: multivariate Gaussian distribution

Definition

Let $\mu \in \mathbb{R}^d$, Σ be a positive definite matrix. We write $X \sim \mathcal{N}(\mu, \Sigma)$ when the Lebesgue density of X is

$$\begin{aligned} x \in \mathbb{R}^d \mapsto & |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \end{aligned}$$

where $|\Sigma|$ is the determinant of Σ. In addition, we have

$$\mathbb{E}X = \mu, \quad \mathbb{V}(X) = \Sigma,$$

where $\mathbb{V}(X)$ is the covariance matrix of X.

Recall the MLEs

Question: what are the MLEs for the expectation and the covariance matrix of a Gaussian sample?

Proposition

Let $\mu^\star \in \mathbb{R}^d$, $\Sigma^\star$ *be a positive definite matrix and* $\{X_1, \ldots, X_n\}$ *be a sample i.i.d. according to* $\mathcal{N}(\mu^\star, \Sigma^\star)$*.*
Then

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^\top$$

are maximum likelihood estimators (MLEs) respectively of $\mu^\star$ *and* $\Sigma^\star$*.*

Proof I

Let $\varphi \colon (x, \mu, \Sigma) \in \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{S} \mapsto |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}$, where $\mathbb{S}$ is the set of positive definite matrices. We aim at maximizing the log-likelihood of $(X_1, \ldots, X_n)$ with respect to μ and Σ, that is solving the convex optimization problem

$$\min_{\mu \in \mathbb{R}^d,\, \Sigma \in \mathbb{S}} \ell_n(\mu, \Sigma) = \sum_{i=1}^n \frac{1}{2}(X_i - \mu)^\top \Sigma^{-1}(X_i - \mu) - \frac{n}{2} \log\left(|\Sigma^{-1}|\right).$$

Since the objective function is differentiable on $\mathbb{R}^d \times \mathbb{S}$, we first solve the problem with respect to μ, leading to the solution $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$.
Then, it remains so solve

$$\min_{\Sigma \in \mathbb{S}} \sum_{i=1}^n \frac{1}{2}(X_i - \hat{\mu})^\top \Sigma^{-1}(X_i - \hat{\mu}) - \frac{n}{2} \log\left(|\Sigma^{-1}|\right).$$

Proof II

Reminding that for any square matrices A and B, $\frac{\partial}{\partial A}\operatorname{tr}(AB) = B^\top$ and $\frac{\partial}{\partial A}\log(|A|) = (A^{-1})^\top$ and setting the gradient of the objective function with respect to Σ^{-1} to 0 leads to the positive definite solution (as soon as $n \geq d$ and $\{X_1, \dots, X_n\}$ are all different almost surely) $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^\top$.

Discriminant analysis

- $(X, Y) \in \mathbb{R}^d \times \{1, \dots C\}$ be a pair of r.v.
- Y is a label characterizing the class of X.
- **Goal:** computing the Bayes classifier when each class $i \in \{1, \dots, C\}$ is normally distributed:
  there exists a positive definite matrix Σ_i and a vector $\mu_i \in \mathbb{R}^d$ such that

$$X | Y = i \sim \mathcal{N}(\mu_i, \Sigma_i).$$

Recall: a Bayes classifier

For multiclasses

$$\forall x \in \mathbb{R}^d : \quad g^\star(x) \in \operatorname{argmax}_{i \in [C]} \mathbb{P}(Y = i | X = x).$$

Bayes classifier for normal distributions

Proposition

Let us assume that each class is normally distributed and let $\pi_i = \mathbb{P}(Y = i)$ *be class prior probabilities, for all* $i \in [C]$. *Then, a Bayes classifier* $g^\star$ *is defined by:* $\forall x \in \mathbb{R}^d$

$$g^\star(x) \in \mathrm{argmax}_{i \in [C]} \log(\pi_i) - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i).$$

Proof: Compute the log-ratio of the conditional probabilities.

The case of two classes

- Only two classes ($C = 2$)
- In this case, a Bayes classifier satisfies

$$g^\star \colon x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1 | X = x) > \mathbb{P}(Y = 2 | X = x) \\ 2 & \text{otherwise.} \end{cases}$$

Linear discriminant analysis (LDA)

- Each class is normally distributed

$$X|Y = i \sim \mathcal{N}(\mu_i, \Sigma), \quad i = 1, 2$$

- With equal covariance

Proposition

Let $\pi_i = \mathbb{P}(Y = i)$ *be class prior probabilities, for* $i \in \{1, 2\}$,

$$h \colon x \in \mathbb{R}^d \mapsto (\mu_1 - \mu_2)^\top \Sigma^{-1} x$$

$$b = \frac{1}{2}(\mu_2^\top \Sigma^{-1} \mu_2 - \mu_1^\top \Sigma^{-1} \mu_1) + \log\left(\frac{\pi_1}{\pi_2}\right).$$

Then, a Bayes classifier is

$$g^\star \colon x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \textit{if } h(x) + b > 0 \\ 2 & \textit{otherwise.} \end{cases}$$

Proof: Straightforward.

LDA

- Note that the function $h(x) + b$ is linear in x.
- This is a linear classifier!

What happens when $\pi_1 = \pi_2$

- if $\pi_1 = \pi_2$, we have:

$$g^\star(x) = 1 \\ \iff (x - \mu_1)^\top \Sigma^{-1} (x - \mu_1) < (x - \mu_2)^\top \Sigma^{-1} (x - \mu_2),$$

- $\pi_1 = \pi_2$ if and only if x is closer to μ_1 than μ_2 with respect to the Mahalanobis distance ruled by Σ.

- This is similar to whitening the data with $\Sigma^{-\frac{1}{2}}$ and considering the Euclidean distance.

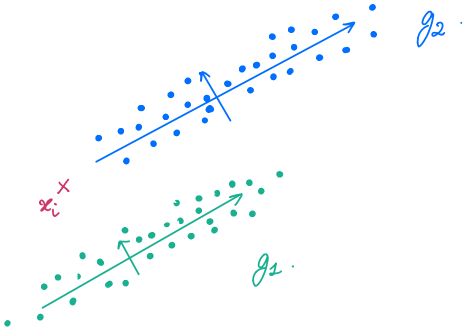


Figure: Here, Point x_i is closer to g_1 in the Euclidean distance while it appears naturally that it belongs to the group of data centered in g_2. The Mahalanobis distance makes it possible to rectify this misbehavior.

Quadratic discriminant analysis (QDA)

- Each class is normally distributed
- But with different covariances

Proposition

Let $\pi_i = \mathbb{P}(Y = i)$ be class prior probabilities, for all $i \in \{1, 2\}$, and let us denote

$$h\colon x \in \mathbb{R}^d \mapsto \frac{1}{2} x^\top (\Sigma_2^{-1} - \Sigma_1^{-1}) x + (\mu_1^\top \Sigma_1^{-1} - \mu_2^\top \Sigma_2^{-1}) x$$

$$b = \frac{1}{2}(\mu_2^\top \Sigma_2^{-1} \mu_2 - \mu_1^\top \Sigma_1^{-1} \mu_1) - \frac{1}{2} \log\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right) + \log\left(\frac{\pi_1}{\pi_2}\right).$$

Then, a Bayes classifier is

$$g^\star\colon x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \textit{if } h(x) + b > 0 \\ 2 & \textit{otherwise.} \end{cases}$$

Proof: Left as an exercise.

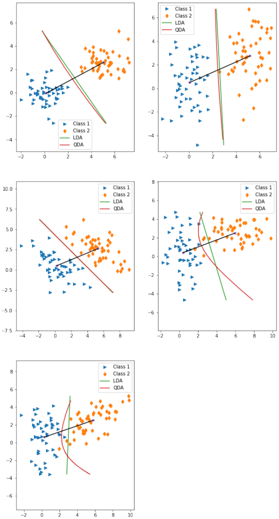


Figure: Comparison of linear discrimant analysis (LDA) and quadratic discriminant analysis (QDA) on different simulated datasets (Gaussian classes with potentially different covariance matrices).

Towards the Rayleigh quotient

Setting

- Suppose that the distribution of $X|Y$ is unimodal
- So the distribution of $X|Y$ can be characterized by the information of
  - (i) its expectation
  - (ii) its variance

Discriminant analysis

Aim: finding a direction $w \in \mathbb{R}^d$ such that the projection of X onto that direction maximizes the variance between classes while minimizing the variances within classes.

Towards the Rayleigh quotient

Def: Rayleigh quotient

$$r = \frac{\text{variance between the classes}}{\text{variance in the classes}}$$

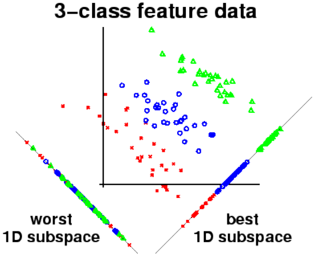


Figure: Subspace with maximal Rayleigh quotient. Courtesy of Quora & Maxime Sangnier.

Towards the Rayleigh quotient

More formally, we are interested in solving

$$\max_{w \in \mathbb{R}^d} r(w) = \frac{\mathbb{V}\left(\mathbb{E}\left(w^\top X | Y\right)\right)}{\mathbb{E}\left(\mathbb{V}\left(w^\top X | Y\right)\right)},$$

- where for any random pair of variables $(U, V) \in \mathbb{R}^2$, $\mathbb{V}\left(U | V\right) = \mathbb{E}\left(\left(U - \mathbb{E}\left(U | V\right)\right)^2 | V\right)$.
- With $\mu = \mathbb{E}X$, $\mu_i = \mathbb{E}(X | Y = i)$, $\Sigma_i = \mathbb{V}(X | Y = i)$ and $\pi_i = \mathbb{P}(Y = i)$, for $i = 1, \dots C$, we have:

$$r(w) = \frac{w^\top \left(\sum_{i=1}^{C} \pi_i (\mu_i - \mu)(\mu_i - \mu)^\top\right) w}{w^\top \left(\sum_{i=1}^{C} \pi_i \Sigma_i\right) w}.$$

Fisher discriminant analysis

- In the case of two classes, $C = 2$.
- With

$$\mu = \pi_1\mu_1 + (1 - \pi_1)\mu_2$$

The Rayleigh quotient becomes

$$\begin{aligned} r(w) & \\ &= \frac{w^\top \left(\pi_1(\mu_1 - \mu)(\mu_1 - \mu)^\top + (1 - \pi_1)(\mu_2 - \mu)(\mu_2 - \mu)^\top\right) w}{w^\top \left(\pi_1\Sigma_1 + (1 - \pi_1)\Sigma_2\right) w} \\ &= \pi_1(1 - \pi_1)\frac{(w^\top\mu_1 - w^\top\mu_2)^2}{w^\top \left(\pi_1\Sigma_1 + (1 - \pi_1)\Sigma_2\right) w}. \end{aligned}$$

Fisher's linear discriminant

Proposition

Let us assume that $C = 2$ and that $\mu_1 \neq \mu_2$. Then

$$\left\{\lambda(\pi_1\Sigma_1 + (1-\pi_1)\Sigma_2)^{-1}(\mu_1 - \mu_2), \lambda \in \mathbb{R}\backslash\{0\}\right\} \subset \operatorname{argmax}_{w \in \mathbb{R}^d} r(w).$$

Proof I

First, let us remark that

$$r(w) = \pi_1(1-\pi_1)\frac{w^\top(\mu_1-\mu_2)(\mu_1-\mu_2)^\top w}{w^\top\left(\pi_1\Sigma_1+(1-\pi_1)\Sigma_2\right)w}.$$

Moreover, since r is differentiable everywhere except in 0, we have that if r attains an extremum in $w \neq 0$, then, necessarily $\nabla r(w) = 0$. By the chain rule, this means

$$\begin{aligned}&[w^\top\left(\pi_1\Sigma_1+(1-\pi_1)\Sigma_2\right)w](\mu_1-\mu_2)(\mu_1-\mu_2)^\top w\\ &\qquad - [w^\top(\mu_1-\mu_2)(\mu_1-\mu_2)^\top w]\left(\pi_1\Sigma_1+(1-\pi_1)\Sigma_2\right)w = 0,\end{aligned}$$

that is, when reorganizing,

$$\begin{aligned}&[w^\top\left(\pi_1\Sigma_1+(1-\pi_1)\Sigma_2\right)w(\mu_1-\mu_2)^\top w](\mu_1-\mu_2)\\ &\qquad = [(\mu_1-\mu_2)^\top w]^2\left(\pi_1\Sigma_1+(1-\pi_1)\Sigma_2\right)w,\end{aligned}$$

Proof II

where the terms in brackets in the left and right hand side are scalar. Assuming that w is not orthogonal to $\mu_1 - \mu_2$ (that is the scalar in the right hand side is nonzero), this means that, in any case:

$$w \propto (\pi_1 \Sigma_1 + (1 - \pi_1)\Sigma_2)^{-1} (\mu_1 - \mu_2).$$

Recapping, a nonzero extremum of r is either orthogonal to $\mu_1 - \mu_2$ or satisfies the previous relation.
On the one hand, considering w orthogonal to $\mu_1 - \mu_2$ leads to $r(w) = 0$ (which is the minimum of r). On the other hand, all nonzero vectors satisfying the previous relation have the same value of Rayleigh quotient, which is not null. Thus, vectors proportional to $(\pi_1 \Sigma_1 + (1 - \pi_1)\Sigma_2)^{-1} (\mu_1 - \mu_2)$ are maxima.

Concluding on Fisher's discriminant analysis

Remark (Fisher $\leftrightsquigarrow$ LDA)

When covariance matrices are equal, Fisher's discriminant direction is the same as the one in linear discriminant analysis (LDA).

Given that direction, projection of X is given by:

$$h(X) = w^\top X.$$

Moreover, an intercept b can be defined by:

$$b \in \operatorname{argmin}_{a \in \mathbb{R}} \mathbb{P}(Y \neq g_a(X)),$$

where

$$g_a \colon x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \text{if } h(x) + a > 0 \\ 2 & \text{otherwise.} \end{cases}$$

Concluding on Fisher's discriminant analysis

Let us remark that, in its empirical version (that is replacing expected values by their means computed with the sample $\{(X_i, Y_i)\}_{1 \leq i \leq n}$), an intercept can be defined by

$$b \in \mathrm{argmin}_{a \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq g_a(X_i)},$$

where $a \in \mathbb{R} \mapsto \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq g_a(X_i))}$ is a piecewise constant function, for which the steps are at $\{-h(X_1), \dots, -h(X_n)\}$. This means that only n values have to be evaluated to determine an empirical threshold b.

Summary

1. Mathematical framework

2. Discriminant analysis
   - The multivariate normal distribution
   - Bayes classifier for multivariate normal distributions
   - Rayleigh quotient
   - Fisher discriminant analysis

3. Logistic regression

Recall the linear model

- In regression with $\mathcal{X} = \mathbb{R}^d$, the linear model is the parametric reference model.
- This model makes the assumption that the regression function is linear:

  $$m^\star(x) = \mathbb{E}[Y_i | X_i = x] = \beta_1^\star x_1 + \ldots + \beta_d^\star x_d,$$

  or equivalently that for $1 \leqslant i \leqslant n$

  $$Y_i = \beta_1 X_{i,1} + \ldots + \beta_d X_{i,d} + \varepsilon_i = X_i^T \beta + \varepsilon_i,$$

  with

  $$\mathbb{E}[\varepsilon_i | X_i = x] = 0 \quad \text{and} \quad \mathbb{V}[\varepsilon_i | X_i = x] = \sigma^2.$$

- Here, estimating $m^\star$ is equivalent to estimate $\beta^\star \in \mathbb{R}^d$. $m^\star$ is characterized by the parameter $\beta^\star \in \mathbb{R}^d$ (finite dimension) $\Rightarrow$ parametric model.

ERM with quadratic loss

- The least squares estimates, i.e. the optimal solution of

$$\min_{\beta \in \mathbb{R}^d} \widehat{\mathcal{R}}_n(\beta) = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 = \min_{\beta \in \mathbb{R}^d} \frac{1}{n} \| Y - X\beta \|_2^2$$

with $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times d}$ is given by

$$\widehat{\beta}^{LS} = (X^T X)^{-1} X^T Y.$$

- The regression function $m^\star$ is estimated by

$$\widehat{m}_n^{LS}(x) = \hat{\beta}_1^{LS} x_1 + \ldots + \hat{\beta}_d^{LS} x_d.$$

- Under some technical assumptions (see lectures of past year)

$$\mathbb{E}[\widehat{\beta}^{LS}] = \beta^\star \quad \text{and} \quad \mathbb{V}(\widehat{\beta}^{LS}) = (X^T X)^{-1} \sigma^2.$$

- We deduce that

$$\mathbb{E}\left[\| \widehat{\beta} - \beta \|_2^2 \right] = O\left(\frac{1}{n} \right) \quad \text{and} \quad \mathbb{E}\left[(\hat{m}_n^{LS}(x) - m^\star(x))^2 \right] = O\left(\frac{1}{n} \right)$$

Regression and linear model

Remark

- *Least squares estimates achieve the parametric rate $O(1/n)$.*
- *Moreover, if errors terms $(\varepsilon_i)_i$ are Gaussian, we can compute the distribution of the least squares estimates (confidence intervals, test statistics...).*

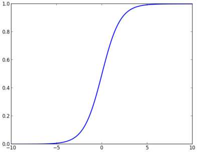
Going beyond: the logistic model

- One of the most widely used classification algorithm.
- Logistic model is the "brother" of the linear model in the context of binary classification ($\mathcal{Y} = \{-1, 1\}$).
- We want to explain the label Y based on X, we want to "regress" Y on X.
- It models the distribution of $Y|X$. For $y \in \{-1, 1\}$

$$\mathbb{P}\left(Y = 1|X = x\right) = \sigma\left(x^T w + b\right)$$

where $w \in \mathbb{R}^d$ is a vector of model weights and $b \in \mathbb{R}$ is the intercept, and where σ is the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$



Some comments on the logistic model

- The sigmoid choice really is a choice. It is a modelling choice.
- It's a way to map $\mathbb{R} \to [0,1]$ (we want to model a probability).
- We could also consider

  $$\mathbb{P}\left(Y = 1 | X = x\right) = F\left(x^T w + b\right)$$

  for any distribution function F.
- Another popular choice is the Gaussian distribution

  $$F(z) = \mathbb{P}(\mathcal{N}(0,1) \leqslant z),$$

  which leads to another loss called probit.

The logistic model

- In the case of the sigmoid, one has

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

- However, the sigmoid choice has the following nice interpretation: an easy computation leads to

$$\log\left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = -1|X = x)}\right) = x^T w + b.$$

- This quantity is called the log-odd ratio.

The logistic model

- Therefore, this model makes the assumption that (the logit transformation of) the probability $p(x) = \mathbb{P}(Y = 1|X = x)$ is linear:

$$\mathrm{logit}(p(x)) := \log\left(\frac{p(x)}{1 - p(x)}\right) = x^T w + b.$$

- Note that

$$\mathbb{P}(Y = 1|X = x) \geqslant \mathbb{P}(Y = -1|X = x)$$

<u>**iff**</u>

$$x^T w + b \geqslant 0.$$

This is a linear classification rule, linear w.r.t. the considered features x!

The logistic "regression"

Theorem

Let us consider that $C = 2$ *and that the logit-transformation is linear with parameters* $(b^\star, w^\star)$*. Let* $f^\star \colon x \in \mathbb{R}^d \mapsto b^\star + (w^\star)^\top x$*. Then* $f^\star$ *is a minimizer of the risk functional* $f \mapsto \mathbb{E}\left[\log\left(1 + \exp(-Yf(X))\right)\right]$ *over all affine functions and*

$$g^\star \colon x \in \mathbb{R}^d \mapsto \begin{cases} +1 & \textit{if } f^\star(x) > 0 \\ -1 & \textit{otherwise} \end{cases}$$

is a Bayes classifier.

Proof I

First, let us remark that $g^\star$, as previously defined, is a Bayes classifier, since $f^\star(x) > 0$ if and only if $\mathbb{P}(Y = +1|X = x) > \mathbb{P}(Y = -1|X = x)$.
Second, without loss of generality, we consider that b vanishes and we denote $\ell \colon (x, y, w) \mapsto \log\left(1 + \exp(-yw^\top x)\right)$. Then, $\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}$,

$$\nabla_w \ell(x, y, w) = -\frac{y}{1 + \exp(yw^\top x)} x = -\frac{y \exp(-yw^\top x)}{1 + \exp(-yw^\top x)} x.$$

Proof II

Thus, since $\nabla_w \ell$ is Lebesgue-measurable, we can switch derivative and integral, leading to:

$$
\begin{aligned}
&\nabla_w \mathbb{E}(\ell(X, Y, w)|X) \\
&\quad = \mathbb{P}(Y = +1|X)\nabla_w \psi(X, 1, w) + \mathbb{P}(Y = -1|X)\nabla_w \psi(X, -1, w) \\
&\quad = -\frac{\exp((w^\star)^\top x)}{1 + \exp((w^\star)^\top x)} \frac{1}{1 + \exp(w^\top X)} X \\
&\qquad + \frac{1}{1 + \exp((w^\star)^\top x)} \frac{\exp(w^\top X)}{1 + \exp(w^\top X)} X.
\end{aligned}
$$

Thus, $\nabla_w \mathbb{E}(\ell(X, Y, w^\star)|X) = 0$ and $\nabla_w \mathbb{E}(\ell(X, Y, w^\star)) = 0$. Since ℓ is convex in w, this proves that $w^\star$ is a minimizer of $\mathbb{E}(\ell(X, Y, \cdot))$.

Estimation of w and b

- We have a model for $Y|X$
- Data (x_i, y_i) is assumed i.i.d with the same distribution as (X, Y)
- Compute estimators $\hat{w}$ and $\hat{b}$ by maximum likelihood estimation
- Or equivalently, minimize the minus log-likelihood.
- More generally, when a model is used

  $$\text{Goodness-of-fit} = \text{-log likelihood}$$

  log is used mainly since averages are easier to study (and compute) than products

The logistic "regression"

By introducing the logistic loss function

$$\ell(y, y') = \log(1 + e^{-yy'}),$$

then

$$\hat{w}, \hat{b} \in \operatorname{argmin}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i^T w + b).$$

- It is a convex and smooth problem
- Many ways to find an approximate minimizer
- Efficient convex optimization algorithms (more on that later)

But...

Remark

However, if there exists a separating hyperplane for $(X_i, Y_i)_{1 \leqslant i \leqslant n}$ *i.e. if there exists* (b_0, w_0) *such that*

$$\forall i = 1, \ldots, n, \qquad Y_i(w_0^T X_i + b_0) > 0,$$

then there is no minimizer of the negative log-likelihood! Convince yourself!

Summary: LDA/QDA/Logistic reg

Logistic regression directly models the parameter of the distribution of $Y|X = x$

- The logit transformation of the probability $p(x) = \mathbb{P}\left(Y = 1|X = x\right)$ is linear:

$$\mathrm{logit}(p(x)) := \log\left(\frac{p(x)}{1 - p(x)}\right) = x^T w + b.$$

Linear discriminant analysis do the opposite.

- It models the distributions of $X|Y = j$ for $j = 1, ..., C$ by Gaussian distributions $f_j(x)$,
- The posterior distribution $Y|X = x$ can be computed with Bayes formula:

$$\mathbb{P}\left(Y = j|X = x\right) = \frac{\pi_j f_j(x)}{\sum_\ell^C \pi_\ell f_\ell(x)},$$

with $\pi_j = \mathbb{P}(Y = j)$.

Summary: LDA/QDA/Logistic reg

- Classification rule: we choose the group which maximizes these probabilities $\hat{g}(x) = k$ if and only if $\mathbb{P}(Y = k|X = x) \geqslant \mathbb{P}(Y = j|X = x), \forall j \neq k$.
- Boundary between 2 groups: set of points x such that

$$\mathbb{P}(Y = k|X = x) = \mathbb{P}(Y = j|X = x)$$

i.e.

$$\begin{aligned} 0 &= \log\left(\frac{\mathbb{P}(Y = k|X = x)}{\mathbb{P}(Y = j|X = x)}\right) = \log\frac{f_k(x)}{f_j(x)} + \log\frac{\pi_k}{\pi_j} \\ &= \log\frac{\pi_k}{\pi_j} + \frac{1}{2}(\mu_k + \mu_j)^T\Sigma^{-1}(\mu_k - \mu_j) + x^T\Sigma^{-1}(\mu_k - \mu_j) \end{aligned}$$

which is linear in x!

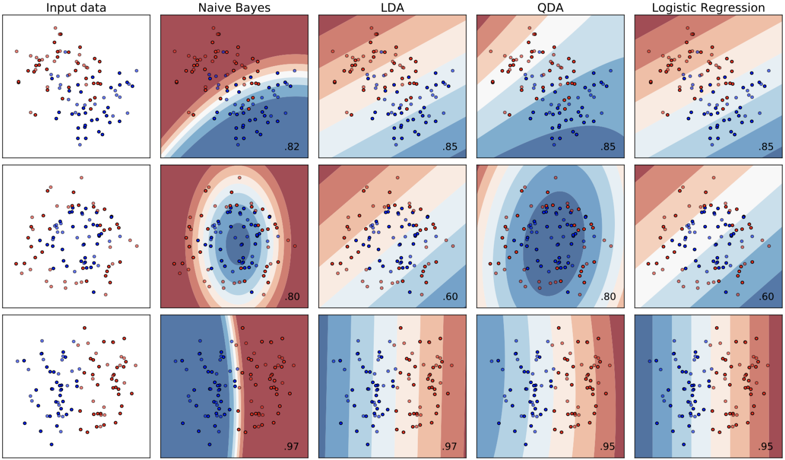


Figure: Comparison between LDA, QDA and logistic regression