Mathématiques pour l'IA 1
Bayesian Decision Theory, Discriminant Analysis

Serge Iovleff

November 20, 2024

« [On the Gaussian curve] *Experimentalists think that it is a mathematical theorem while the mathematicians believe it to be an experimental fact.* »

Gabriel Lippmann

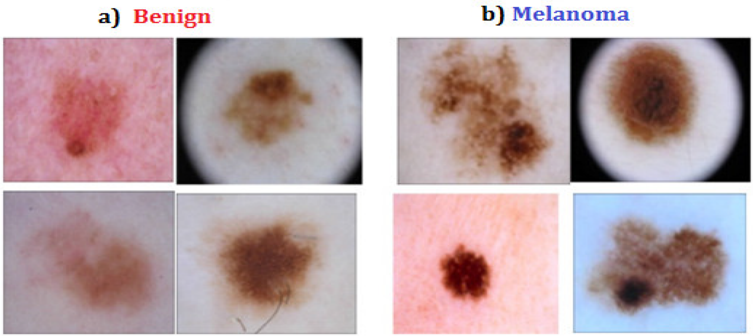
Outline

Statistical Methods

- Statistical methods in machine learning all have in common that they assume that the process that “generates” the data is governed by the rules of probability
- The data is understood to be a set of random samples from some underlying probability distribution
- First part of this talk will be all about probabilistic models. In the second part, and other talks, the use of probability will sometimes be much less explicit
- Nonetheless, the basic assumption about how the data is generated is always there, even if you don’t see a single probability distribution anywhere

Cancer detection

Two classes decision problem



$\Rightarrow$ classify a new image so that the probability of a wrong classification is minimized

Class conditional probabilities

- Probability of making an observation $\mathbf{x}$ knowing that it comes from some class $\mathcal{C}_k$
- Here $\mathbf{x}$ is often a feature vector, which measures/describes properties of the data. E.g.: number of black pixels, height-width ratio, ...

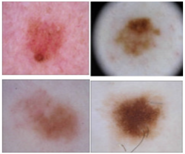


Figure: "benign"

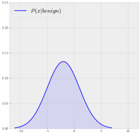


Figure: Distribution of $\mathbf{x}$ conditionally to "benign"

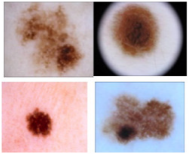


Figure: "malign"

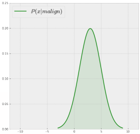


Figure: Distribution of $\mathbf{x}$ conditionally to "malign"

Class conditional probabilities

- Example, we have an observation $\mathbf{x} = -2$

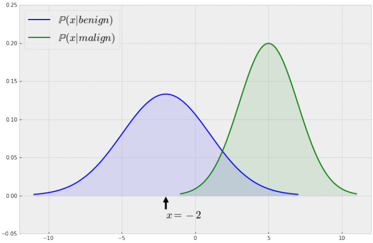


Figure: $\mathbf{x} = -2$ should produce “benign”

- How do we decide which class the data point belongs to?
- Here, we should decide for class “benign”

Class conditional probabilities

- Example, we have an observation $\mathbf{x} = 6$

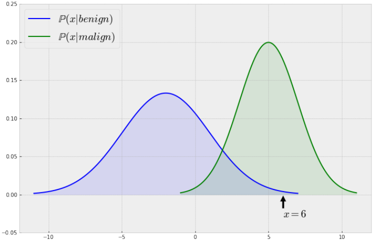


Figure: $\mathbf{x} = 6$ should produce “malign”

- How do we decide which class the data point belongs to?
- Here, we should decide for class “malign”

Class conditional probabilities

- Example, we have an observation $\mathbf{x} = 2$

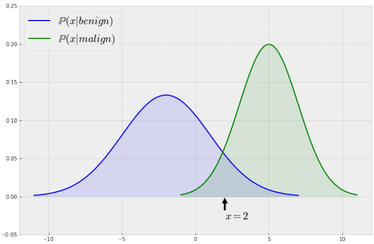


Figure: $\mathbf{x} = 2$ should produce ?

- How do we decide which class the data point belongs to?

Class Priors

- The a priori probability of a data point belonging to a particular class is called the class prior
- Invasive melanomas account for about 1% of all skin cancer cases[1]
- What are $p(\text{“}malign''\text{)}$ and $p(\text{“}benign'')$?

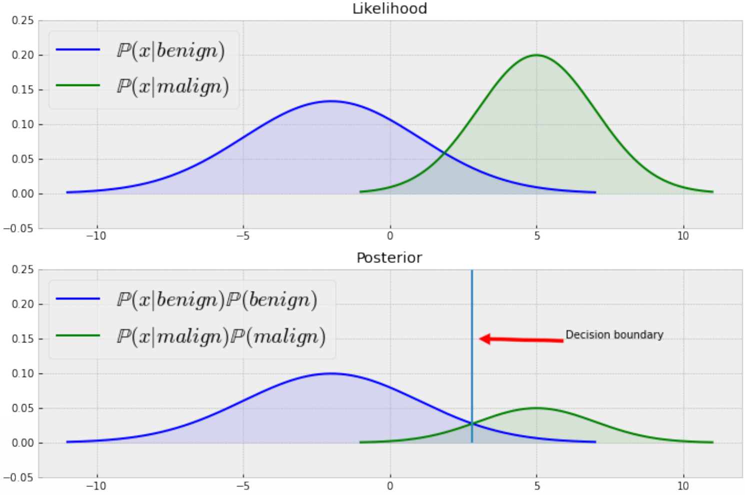
$$
\begin{aligned}
C_1 = \text{“}malign'' \qquad p(C_1) &= 0.01 \\
C_2 = \text{“}benign'' \qquad p(C_2) &= 0.99
\end{aligned}
$$

$$\sum_k p(C_k) = 1$$

- How do we decide which class the data point belongs to?
- If $p(C_1) = 0.01$ and $p(C_2) = 0.99$, we should decide for “benign”.

[1] but they account for over 75% of skin cancer deaths

Bayesian Decision Theory



Bayesian Decision Theory

- Bayes formula

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}.$$

- We want to find the a-posteriori probability (posterior) of the class C_k given the observation (feature) $\mathbf{x}$

$$\mathrm{p}(C_k \mid \mathbf{x}) = \frac{\mathrm{p}(\mathbf{x} \mid C_k)\mathrm{p}(C_k)}{\mathrm{p}(\mathbf{x})}.$$

  - $\mathrm{p}(C_k) = \pi_k$ is the *prior* probability
  - $\mathrm{p}(\mathbf{x} \mid C_k)$ is the class-conditional probability (likelihood)
  - $\mathrm{p}(C_k \mid \mathbf{x})$ is the class *posterior* probability
  - $\mathrm{p}(\mathbf{x})$ is the normalization term

- We classify a new point according to which density is highest. When the priors are different, we take them into account as well

Bayesian Decision Theory

- Decision rule: decide C_1 if $\mathrm{p}(C_1 | \mathbf{x}) > \mathrm{p}(C_2 | \mathbf{x})$
- Equivalent to

$$\frac{\mathrm{p}(\mathbf{x} | C_1)\,\mathrm{p}(C_1)}{\mathrm{p}(\mathbf{x})} > \frac{\mathrm{p}(\mathbf{x} | C_2)\,\mathrm{p}(C_2)}{\mathrm{p}(\mathbf{x})}$$

$$\iff \frac{\mathrm{p}(\mathbf{x} | C_1)}{\mathrm{p}(\mathbf{x} | C_2)} > \frac{\mathrm{p}(C_2)}{\mathrm{p}(C_1)}$$

- A classifier obeying this rule is called a Bayes Optimal Classifier
- Generalization to more than 2 classes:
  - Decide for class k iff it has the highest a-posteriori probability

$$\mathrm{p}(C_k | \mathbf{x}) > \mathrm{p}(C_j | \mathbf{x}), \qquad \forall j \neq k$$

  - Equivalent to

$$\mathrm{p}(\mathbf{x} | C_k)\,\mathrm{p}(C_k) > \mathrm{p}(\mathbf{x} | C_j)\,\mathrm{p}(C_j), \qquad \forall j \neq k$$

- Decision Region: R_1, R_2,... form a partition of the predictor space $\mathcal{X}$

Risk Minimization

- So far, we have tried to minimize the misclassification rate
- There are many cases when not every misclassification is equally bad
- Smoke detector
  - If there is a fire, we need to be very sure that we classify it as such
  - If there is no fire, it is ok to occasionally have a false alarm
- Medical diagnosis
  - If the patient is sick, we need to be very sure that we report them as sick
  - If they are healthy, it is ok to classify them as sick and order further testing that may help clarifying this up

Loss Function

- Key idea : we have to construct a loss function in a way that expresses what we want to achieve

$$\text{loss}(\text{decision} = \text{healthy}|\text{patient} = \text{sick}) >>$$
$$\text{loss}(\text{decision} = \text{sick}|\text{patient} = \text{healthy})$$

- Possible decisions: α_i
- True classes: C_j
- Loss function: $\lambda(\alpha_i|C_j)$

  $\Rightarrow$ Measure the loss of deciding α_i when the truth is C_j

Risk Minimization

- The expected loss of a decision is also called the risk of making a decision
- Instead of minimizing the Misclassification Rate

$$
\begin{aligned}
\mathrm{p}(\text{error}) &= \mathrm{p}(\mathbf{x} \in R_1, C_2) + \mathrm{p}(\mathbf{x} \in R_2, C_2) \\
&= \int_{R_1} \mathrm{p}(\mathbf{x} | C_2)\, \mathrm{p}(C_2)\, dx + \int_{R_2} \mathrm{p}(\mathbf{x} | C_1)\, \mathrm{p}(C_1)\, dx
\end{aligned}
$$

- We minimize the Overall Risk

$$
R(\alpha_i|\mathbf{x}) = \sum_j \lambda(\alpha_i|C_j)\mathrm{p}(C_j|\mathbf{x})
$$

- Goal: Create a decision rule so that overall risk is minimized

  $\Rightarrow$ Decide α_1 if $R(\alpha_2|\mathbf{x}) > R(\alpha_1|\mathbf{x})$

Outline

Discriminant Analysis

- Introduction:
  - Model the distribution of $\mathbf{X}$ using Gaussian distribution in each of the classes separately, and then use Bayes theorem to flip things around and obtain $\mathrm{p}(Y|\mathbf{X})$.
  - This leads to linear or quadratic discriminant analysis (LDA or QDA).
- Pros:
  - When the classes are well-separated or n is small, the parameter estimates for the logistic regression (*more on logistic regression later*) model are surprisingly unstable. LDA does not suffer from this problem
  - It is a simple and computationally efficient algorithm
  - Linear Discriminant Analysis (LDA), when we have more than two response classes, provides low-dimensional views of the data.
- Cons:
  - It requires Normal distribution assumption on features/predictors.
  - It assumes that the data is linearly separable
  - It may not perform well in high-dimensional feature spaces

LDA explained when $p = 1$

- The Gaussian density has the form

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

  Here μ_k is the mean, and σ_k^2 the variance (in class k).
- Plugging this into Bayes formula, we get a rather complex expression for $p_k(x) = \mathbb{P}(Y = k \,|\, X = x)$:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma_l} e^{-\frac{(x-\mu_l)^2}{2\sigma_l^2}}}$$

- Happily, there are simplifications and cancellations.

Discriminant functions

Assume $\sigma_k = \sigma$ for all k.

- To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest.
- Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest discriminant score:

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

- Note that $\delta_k(x)$ is a linear function of x.
- If there are $K = 2$ classes and $\pi_1 = \pi_2 = 0.5$, then one can see that the decision boundary is at

$$x = \frac{\mu_1 + \mu_2}{2}.$$

1 What is the discriminant score if σ_k are not all equals ? [Hint: QDA]

Estimating the parameters ($p = 1$)

Maximum likelihood estimates are

$$
\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \qquad \text{with } n_k \text{ the number of samples in class } k \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\
\hat{\sigma}_k^2 &= \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 (\text{In class variances}) \\
\hat{\sigma}^2 &= \frac{1}{n} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 = \sum_{k=1}^{K} \hat{\pi}_k \hat{\sigma}_k^2 \quad (\text{Within variance})
\end{aligned}
$$

Linear Discriminant Analysis ($p > 1$)

- When there is more than one variable, we use the multivariate Gaussian distribution given by

$$f\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Sigma}_k\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)^\top \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)\right]$$

- $\boldsymbol{\mu}_k \in \mathbb{R}^p$ the expectation of $\mathbf{x}$ conditional to $Y = k$
- $\boldsymbol{\Sigma}_k$ the **covariance** matrix of $\mathbf{x}$ conditional to $Y = k$
- If $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all k then the discriminant function is:

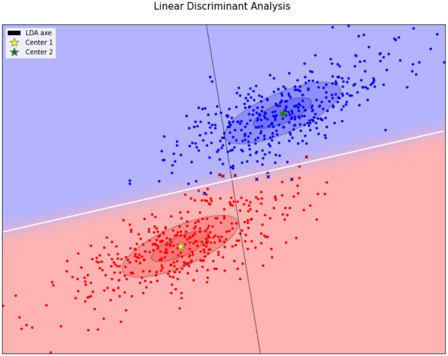
$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k$$

  This is a **linear** function of $\mathbf{x}$!

Outline

FDA versus PCA

- FDA: find a linear combination of features that characterizes or separates two or more classes
- PCA maximizes the variance of projections on the subspace
- FDA maximizes differentiation between classes in subspace



FDA: Discriminant axis

- Covariance between variables:
  - Between-class: S_B calculated by considering that the observations are the centers of gravity of the classes

$$\mathbf{S}_B = \sum_{k=1}^{K} (\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})(\hat{\boldsymbol{\mu}}_k - \hat{\boldsymbol{\mu}})^T$$

  - Within-class: S_W calculated on the initial observations, by centering each class on its center of gravity

$$\mathbf{S}_W = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T$$

  - Total: S calculated on initial observations;

$$\mathbf{S} = \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

  - Huygens relationship $\mathbf{S} = \mathbf{S}_B + \mathbf{S}_W$

Solution

- The first Linear Discriminant projection is find by maximizing the criterion

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

- Solving the generalized eigenvalue problem $\mathbf{S}_B \mathbf{w} = \lambda \mathbf{S}_W \mathbf{w}$ yields the discriminant axis (see PCA)
  - $\mathbf{S}_B$ is the sum of K matrices of rank 1 (or less) and the mean vectors are constrained by a linear relationship

    $\Rightarrow$ $\mathbf{S}_B$ is of rank $K-1$ (or less)

    $\Rightarrow$ if $K < p$, there is at least $K-1$ discriminant axis
  - The matrix $\mathbf{S}_W^{-1} \mathbf{S}_W$ is not symmetric

    $\Rightarrow$ The Discriminant axis are not orthogonal

Outline

Naive Bayes Approach

- Naive Bayes approach is a simple but surprisingly powerful algorithm for predictive modeling.
- If $\mathbf{X} = (X^1, \dots, X^p)$ is a (high) p dimensional vector of features, it may be difficult to find a plausible distribution $\mathbb{P}\left(\mathbf{X} = \mathbf{x} \,|Y = k\,\right)$.
- Using the naive assumption (hence the name) that variables are independent given the class, we can rewrite $\mathbb{P}(\mathbf{X}|y)$ as follows:

$$\mathrm{p}(\mathbf{X}|y) = \prod_{j=1}^{p} \mathrm{p}(X^j|y).$$

- Classification rule become

$$\hat{y} = \arg\max_y \mathrm{p}(y) \prod_{j=1}^{p} \mathrm{p}(x^j|y)$$

Outline

Generative vs. Discriminative

- There are two different views to solve the classification problem
- Generative modelling
  - We model the class-conditional distributions $p(\mathbf{x}|C_2)$ and $p(\mathbf{x}|C_1)$
  - We classify by computing the class posterior using Bayes rule
  - E.g.: Naive Bayes, LDA, QDA
- Discriminative modelling
  - We model the class-posterior directly, e.g. $p(C_1|\mathbf{x})$
  - Consequence: We only care about optimizing the classification rate, and not whether we fit the class-conditional well
  - E.g.: Logistic Regression, Perceptron,...

Probabilistic Discriminative Models

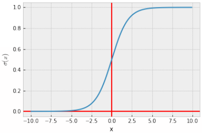
- For now, we will write the class posterior using Bayes' rule

$$
\begin{aligned}
\mathrm{p}(C_1 \mid \mathbf{x}) &= \frac{\mathrm{p}(\mathbf{x} \mid C_1)\,\mathrm{p}(C_1)}{\mathrm{p}(\mathbf{x})} \\
&= \frac{\mathrm{p}(\mathbf{x} \mid C_1)\,\mathrm{p}(C_1)}{\mathrm{p}(\mathbf{x} \mid C_1)\,\mathrm{p}(C_1) + \mathrm{p}(\mathbf{x} \mid C_2)\,\mathrm{p}(C_2)} \\
&= \frac{1}{1 + (\mathrm{p}(\mathbf{x} \mid C_2)\,\mathrm{p}(C_2)) / (\mathrm{p}(\mathbf{x} \mid C_1)\,\mathrm{p}(C_1))} \\
&= \frac{1}{1 + \exp(-a(\mathbf{x}))} \quad \rightarrow \text{logistic sigmoid function}
\end{aligned}
$$

with $a(\mathbf{x}) = \log \frac{\mathrm{p}(\mathbf{x} \mid C_1)\mathrm{p}(C_1)}{\mathrm{p}(\mathbf{x} \mid C_2)\mathrm{p}(C_2)}$

- Logistic/Sigmoid function

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$



$\Rightarrow$ Sigmoid: 'S-shaped'

- Squashes real numbers into the $[0, 1]$ interval

Probabilistic Discriminative Models

- Class posterior

$$\mathrm{p}\left(C_1 \mid \mathbf{x}\right) = \sigma(a) \quad \text{with} \quad a = a(\mathbf{x}) = \log \frac{\mathrm{p}\left(\mathbf{x} \mid C_1\right) \mathrm{p}\left(C_1\right)}{\mathrm{p}\left(\mathbf{x} \mid C_2\right) \mathrm{p}\left(C_2\right)}$$

- Logistic Regression
  - Assume that a is given by a linear discriminant function

$$\mathrm{p}\left(C_1 \mid \mathbf{x}\right) = \sigma(\boldsymbol{\beta}^T \mathbf{x} + \beta_0)$$

  - Find $\boldsymbol{\beta}$ and β_0 so that the class-posterior is modeled best
  - When is this an appropriate assumption?
    - When the class conditionals are Gaussian with equal covariance
    - But also for a number of other distributions
    - There exists some independence of the form of the class-conditionals (*Naive Bayes models*)

Logistic Regression

- Model the class posterior as

$$\mathrm{p}\left(C_1 \,|\mathbf{x}\,\right) = \sigma(\boldsymbol{\beta}\mathbf{x} + \beta_0)$$

- Maximize likelihood

  $\Rightarrow$ Data (as always) is i.i.d. and define $y_i = \begin{cases} 0 & \mathbf{x}_i \text{ belongs to } C_1 \\ 1 & \mathbf{x}_i \text{ belongs to } C_2 \end{cases}$

- Likelihood is

$$\begin{aligned} L(\boldsymbol{\beta}, \beta_0|\mathbf{Y}, \mathbf{X}) &= \prod_{i=1}^{n} \mathrm{p}\left(y_i \,|\mathbf{x}_i; \boldsymbol{\beta}, \beta_0\,\right) \\ &= \prod_{i=1}^{n} \mathrm{p}\left(C_2 \,|\mathbf{x}_i; \boldsymbol{\beta}, \beta_0\,\right)^{y_i} \mathrm{p}\left(C_1 \,|\mathbf{x}_i; \boldsymbol{\beta}, \beta_0\,\right)^{1-y_i} \\ &= \prod_{i=1}^{n} \sigma(\boldsymbol{\beta}^T\mathbf{x}_i + \beta_0)^{y_i} (1 - \sigma(\boldsymbol{\beta}^T\mathbf{x}_i + \beta_0))^{1-y_i} \end{aligned}$$

- The log-likelihood is given by

$$l_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log \sigma(\mathbf{x}_i^T\boldsymbol{\beta}) + (1 - y_i) \log(1 - \sigma(\mathbf{x}_i^T\boldsymbol{\beta})).$$

First derivatives

Compute ∇l_n

- Taking the first derivative with respect to β_j, we get

$$\begin{aligned}\frac{\partial l_n}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{y_i}{\sigma(\mathbf{x}_i^T \boldsymbol{\beta})} \sigma'(\mathbf{x}_i^T \boldsymbol{\beta}) x_{ij} - \frac{1 - y_i}{1 - \sigma(\mathbf{x}_i^T b)} \sigma'(\mathbf{x}_i^T \boldsymbol{\beta}) x_{ij} \\ &= \sum_{i=1}^{n} x_{ij} \sigma'(\mathbf{x}_i^T \boldsymbol{\beta}) \left(\frac{y_i}{\sigma(\mathbf{x}_i^T \boldsymbol{\beta})} - \frac{1 - y_i}{1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})} \right) \\ &= \sum_{i=1}^{n} x_{ij} \frac{\sigma'(\mathbf{x}_i^T \boldsymbol{\beta})}{\sigma(\mathbf{x}_i^T \boldsymbol{\beta})(1 - \sigma(\mathbf{x}_i^T \boldsymbol{\beta}))} (y_i - \sigma(\mathbf{x}_i^T \boldsymbol{\beta})).\end{aligned}$$

- $\sigma' = \sigma(1 - \sigma)$ [Proove it !] which means this simplifies to

$$\frac{\partial l_n}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} (y_i - \sigma(\mathbf{x}_i^T \boldsymbol{\beta}))$$

- So

$$\nabla l_n(\boldsymbol{\beta}) = \mathbf{X}^T (\mathbf{y} - \hat{\mathbf{y}}).$$

Second derivatives

Compute $\nabla^2 l_n$.

- Furthermore

$$\frac{\partial^2 l_n}{\partial\beta_k \partial\beta_j} = -\sum_{i=1}^n x_{ij} \frac{\partial}{\partial\beta_k} \sigma(\mathbf{x}_i^T \boldsymbol{\beta}) = -\sum_i x_{ij} x_{ik} \left[\sigma(\mathbf{x}_i^T \boldsymbol{\beta})(1-\sigma(\mathbf{x}_i^T \boldsymbol{\beta}))\right].$$

- Let

$$\begin{aligned} W &= \mathrm{diag}\left(\sigma(\mathbf{x}_1^T \boldsymbol{\beta})(1-\sigma(\mathbf{x}_1^T \boldsymbol{\beta})), \ldots, \sigma(\mathbf{x}_n^T \boldsymbol{\beta})(1-\sigma(\mathbf{x}_n^T \boldsymbol{\beta}))\right) \\ &= \mathrm{diag}\left(\hat{y}_1(1-\hat{y}_1), \ldots, \hat{y}_n(1-\hat{y}_n)\right). \end{aligned}$$

- Then we have

$$\nabla^2 l_n = -\mathbf{X}^T W \mathbf{X}.$$

- As $\hat{y}_i \in (0,1)$, $-\mathbf{X}^T W \mathbf{X}$ will always be strictly negative definite, although numerically if $\hat{y}_i$ gets too close to 0 or 1 then we may have weights round to 0 which can make H negative semidefinite and therefore computationally singular.

Newton-Raphson algorithm

Use iterative weighted least squares regression.

- Create the working response $\mathbf{z} = W^{-1}(\mathbf{y} - \hat{\mathbf{y}})$ and note that

$$\nabla l_n = \mathbf{X}^T(y - \hat{y}) = \mathbf{X}^T W \mathbf{z}.$$

- All together this means that we can optimize the log likelihood by iterating

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} + (\mathbf{X}^T W^{(k)} \mathbf{X})^{-1} \mathbf{X}^T W^{(k)} \mathbf{z}^{(k)}$$

- Remark: $(\mathbf{X}^T W^{(k)} \mathbf{X})^{-1} \mathbf{X}^T W^{(k)} \mathbf{z}^{(k)}$ is exactly $\hat{\boldsymbol{\beta}}$ for a weighted least squares regression of $\mathbf{z}^{(k)}$ on $\mathbf{X}$.