Linear Classification 2 – Logistic Regression

Julian Karch

Universiteit Leiden
The Netherlands

Topics

- Discriminative vs Generative Classification Part 1
- Logistic Regression
- Discriminative vs Generative Classification Part 2

Generative vs Discriminative Part 1

Universiteit Leiden
The Netherlands

Last Week: Generative Classification

- Estimated the full joint distribution $p(x, y) = p(x|y)p(y)$
- For each new sample, classify to the highest posterior probability $p(y|x)$
- Called generative classification
- **Generative classification methods essentially differ by their model for the likelihood $\boldsymbol{p(x|y)}$ and the estimation technique**
- Rather indirect approach of constructing a classifier

Discriminative Classification

- Just find a classifier C^* *directly* that performs best according to some metric
- Formally, C^* should minimize expected prediction error

$$EPE(C^*) = \mathbb{E}_{X,Y}[L(Y, C^*(X))]$$

- Once we know the true joint distribution, we know the optimal classifer C^* (Bayes Classifier from last week) but not the other way around
- This reveals that recovering the full joint distribution is harder than finding the optimal classifer C^*.

Discriminative Classification

- More often than not, we use the 0-1 loss.

$$L_{01}(y, C(x)) = \begin{cases} 1 & \text{if } y \neq C(x) \\ 0 & \text{if } y = C(x) \end{cases}$$

In this case the, EPE is just the misclassification rate

- Of course, we don't know the joint distribution, so we have to fall back to *empirical risk (ER) minimization* using the training set of size N

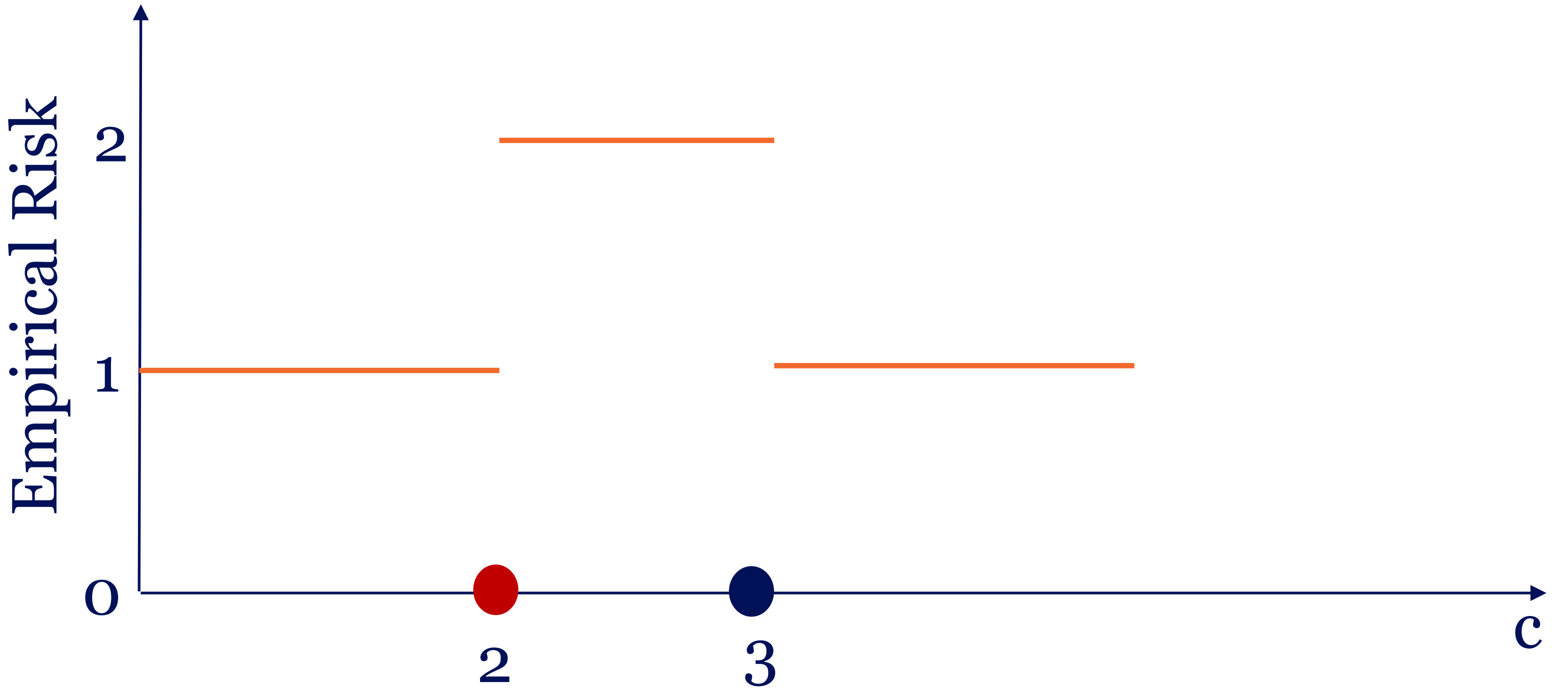
$$\hat{C} = \text{argmin}_{C \in \mathcal{C}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, C(x_i))$$

- *For 0-1 loss: choose the classifier* $\hat{C}$ that minimizes the misclassification rate on the training set

Example Empirical Risk

$$C(x) = \begin{cases} \text{blue}, & x<c \\ \text{red}, & x \geq c \end{cases}$$

- $C(x)$ = blue if x<c, otherwise red
- If c<2, not x<c for both examples
  - → both classified red, ER=1
- If 2<c<3, both classified incorrectly, ER=2
- If c>3, x<c for both → both classified blue, ER=1



Question Emperical Risk

$$C(x) = \begin{cases} 0, & x \leq c \\ 1, & x > c \end{cases}$$

- Classifier is again $C(x) = 0$ if $x \leq c$, otherwise 1
- Training set is: $x_1 = 3, y_1 = 0; x_2 = 5, y_2 = 1$
- Denote all values of c for which this classifier has minimal empirical risk

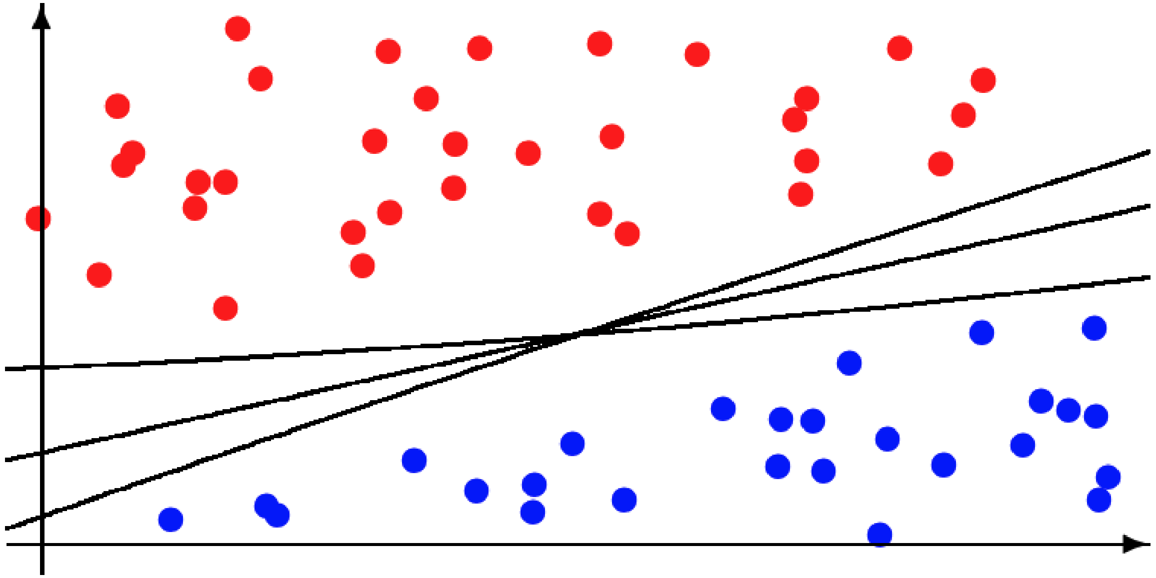
$3 \leq c < 5$

Problems 0-1 Loss

Loss func discrete

Core problem: discontinuous ER

- Finding function that minimizes it is often computationally infeasible (NP-hard)
- There is often no unique minimum



Surrogate Losses

- We care about L_{01} loss but this is impossible / hard to minimize
- Idea: find surrogate loss that is easy to minimize and similar to 0-1 loss in the sense of
  - 0-1 loss large $\Leftrightarrow$ surrogate loss large
  - 0-1 loss small $\Leftrightarrow$ surrogate loss small
- Two popular surrogate losses
  - Logistic loss (today)
  - Hinge loss (support vector machines lecture) SVM
- **Discriminative classification methods are essentially defined by the set of classifiers C they consider and the surrogate loss they employ**

Logistic Loss First Step

- Instead of considering classifiers that return classes (0,1), we consider classifiers that return the posterior probability $P(Y = 1|X) = f(x)$
- We can easily return classes via

$$C(x) = \begin{cases} 1 & \text{if } f(x) >= 0.5 \\ 0 & \text{if } f(x) < 0.5) \end{cases}$$

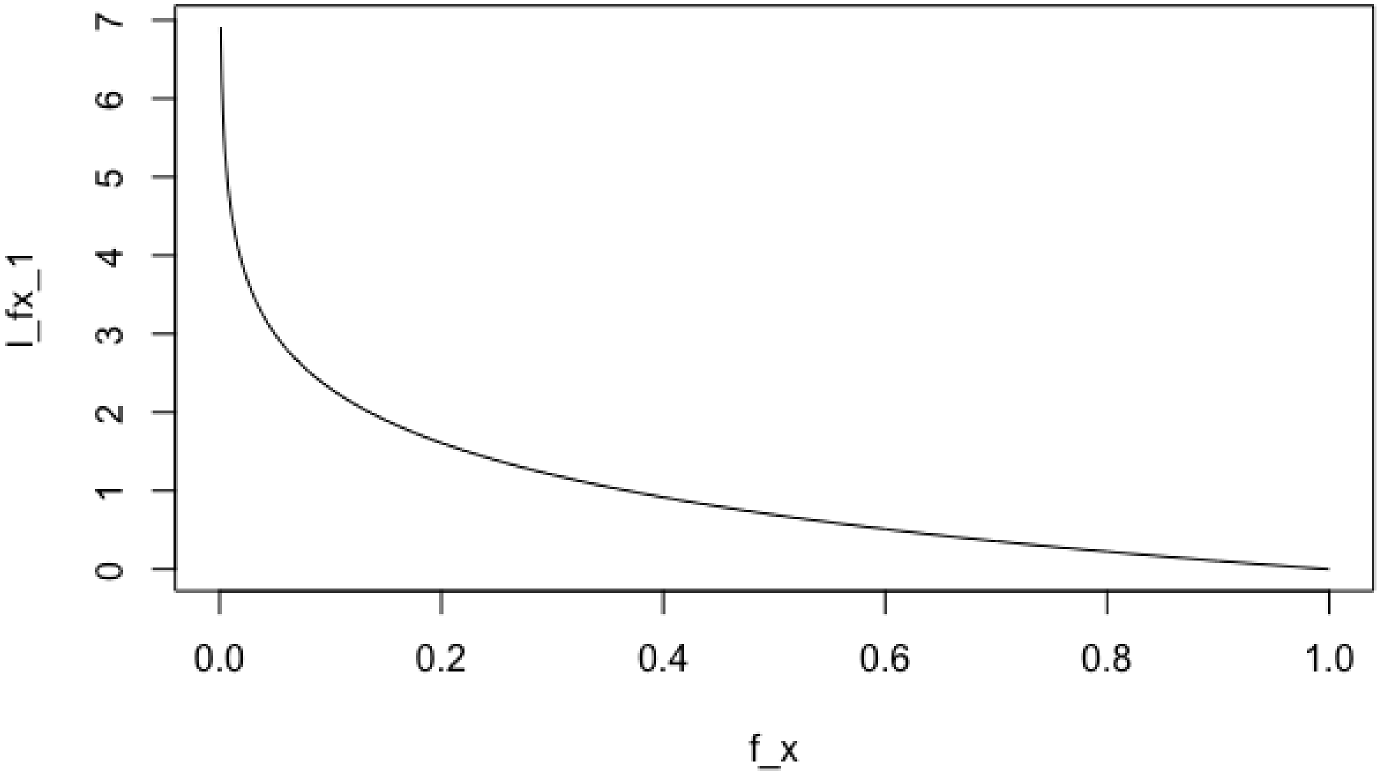
Logistic Loss Motivation

Question:

1. What is f(x) $= P(Y = 1|X = x_i)$ in the best case if $y_i = 1$? 1
2. What is f(x) $= P(Y = 1|X = x_i)$ in the worst case if $y_i = 1$ 0

- For the best case, we want loss 0 → For $L(f(x), 1)$, $L(1,1) = 0$
- For the worst case, we want loss "infinite" → $L(0,1) = \infty$
- In between the loss should increase according to a convex and continuous function

Logistic Loss Intuition



Logistic Loss Motivation

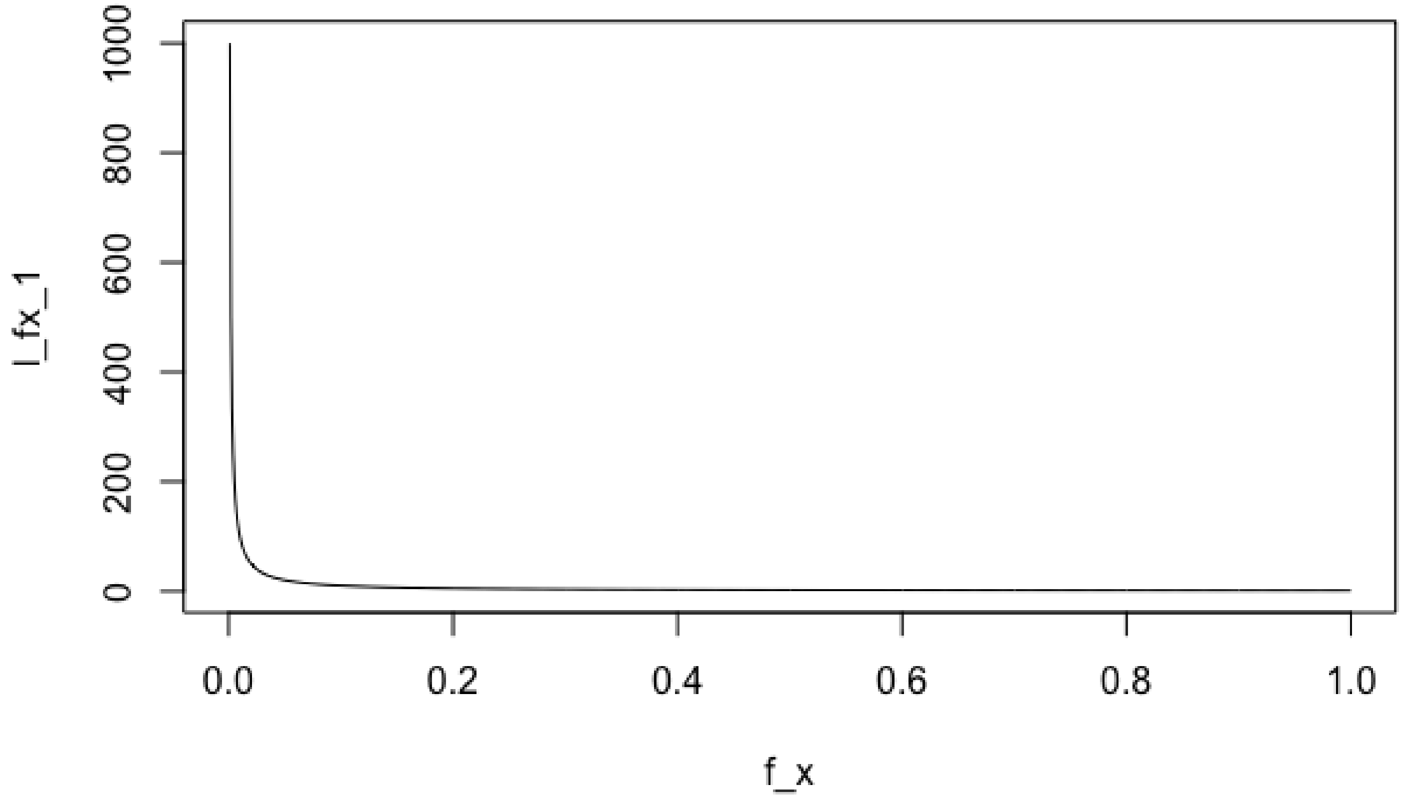
Question:

1. What is f(x) $= P(Y = 1|X = x_i)$ in the best case if $y_i = 0$?
2. What is f(x) $= P(Y = 1|X = x_i)$ in the worst case if $y_i = 0$?

Logistic Loss

$$L_{\log}(y, f(x)) = \begin{cases} -\log(f(x)) & \text{if } y = 1 \\ -\log(1 - f(x)) & \text{if } y = 0 \end{cases}$$

Why not a different loss, like $(1/x)$?



Logistic Loss Has Probabilistic Motivation

- We want to find $f(x)$ that maximizes the likelihood of the training data

$$L(f) = \prod_{i=1}^{N} \hat{P}(Y = y_i | X = x_i)$$

- Note that

$$\hat{P}(Y = y_i | X = x_i) = \begin{cases} f(x_i) & \text{if } y_i = 1 \\ 1 - f(x_i) & \text{if } y_i = 0 \end{cases}$$

- And that maximizing the likelihood is equivalent to mimizing the negative log likelihood

$$-LL(f) = \sum_{i=1}^{N} -\log(\hat{P}(Y = y_i | X = x_i)) = \sum_{i=1}^{N} L_{\log}(y_i, f(x_i))$$

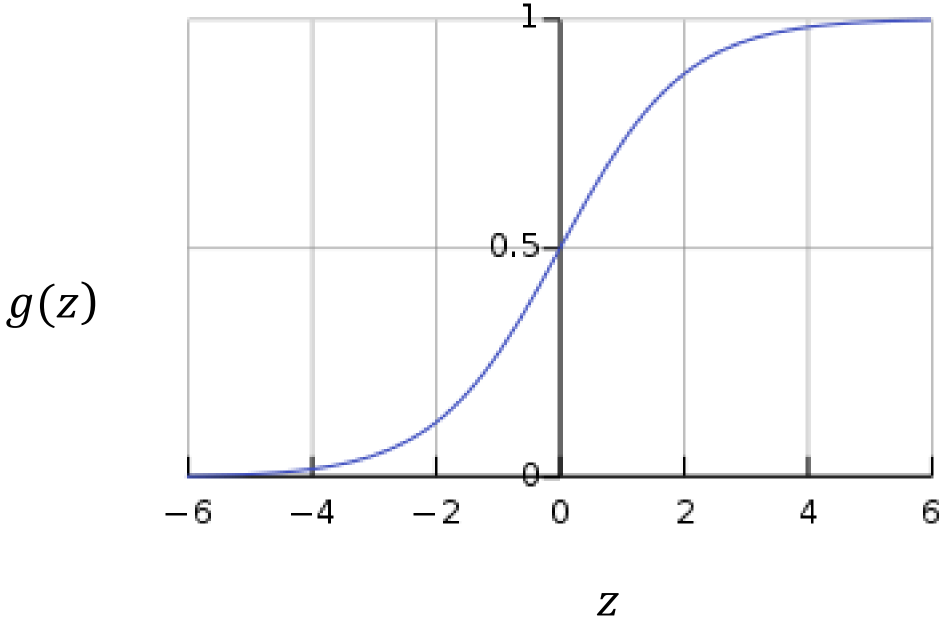
Model for Conditional Probability

- For the problem to become solveable, we need to restrict the set $\mathcal{F}$ of feasible functions, so $f \in \mathcal{F}$
- First idea: Let f be a linear function $f(x;\beta) = \beta^\top x$, as in linear regression
- Problem: f is unbounded but $P(Y = 1|X)$ lies in [0,1]
- Second idea: transform the unbounded linear function using a transformation $g(z)$ such that $g\big(f(x)\big)$ lies in [0,1]
- Especially, we want
  - $\lim_{z\to\infty} g(z) = 1$
  - $\lim_{z\to-\infty} g(z) = 0$

  g(z) to be well behaved (monotonic, continous)

Logistic Function

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



- Note other choice would be possible as well
- Popular example: Probit function, leads to Probit Regression

Logistic Regression Summary Training

We are looking for

$$\text{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{N} L_{\log}(y_i, f(x_i))$$

With

$$\mathcal{F} = \left\{ \frac{1}{1 + e^{-\beta^\top x}} : \beta \in \mathbb{R}^{p+1} \right\}$$

Or

$$\underline{\text{argmin}_{\beta \in \mathbb{R}^{p+1}} \sum_{i=1}^{N} L_{\log}(y_i, \frac{1}{1 + e^{-\beta^\top x}})}$$

Logistic Regression Summary Classification

$$C(x) = \begin{cases} 1 & \text{if } f(x) >= 0.5 \\ 0 & \text{if } f(x) < 0.5) \end{cases}$$

$$C(x) = \begin{cases} 1 & \text{if } \beta^\top x >= 0 \\ 0 & \text{if } \beta^\top x < 0 \end{cases}$$

Logistic Regression Last Question

- Is logistic regression a linear classifier?

Finding The Weights

- In linear regression, we can find the regression weights β "directly"

- In logistic regression, we have to rely on iterative optimization procedures

- Details are beyond the scope of this course

Disclaimer

- Using logistic regression in an inferential fashion, that is, interpreting p-values to make conclussions about population values requires that you have strong reasons to believe that its assumptions are fullfilled.
- Core assumption:

$$P(Y = 1|X = x) = \frac{1}{1 + e^{-\beta^\top x}}$$

Generative vs Discriminative Part 2

Universiteit Leiden
The Netherlands

Implied Discriminative Model LDA 1

$$
\begin{aligned}
P(Y=1|X=x) &= \frac{P(x|Y=1)P(Y=1)}{P(x|Y=1)P(Y=1)+P(x|Y=2)P(Y=2)} \\
&\hat{=} \frac{\hat{f}_1(x)\hat{\pi}_1}{\hat{f}_1(x)\hat{\pi}_1+\hat{f}_2(x)\hat{\pi}_2} \\
&= \frac{1}{1+\frac{\hat{f}_2(x)\hat{\pi}_2}{\hat{f}_1(x)\hat{\pi}_1}}
\end{aligned}
$$

$$\frac{1}{1+e^{-\beta^T x}}$$

Implied Discriminate Model LDA 2

$$
\begin{aligned}
\frac{\hat{f}_2(x)\hat{\pi}_2}{\hat{f}_1(x)\hat{\pi}_1} &= \frac{(2\pi)^{-\frac{p}{2}} \det\left(\hat{\Sigma}^{-\frac{1}{2}}\right) \exp\left(-\frac{1}{2}(x-\hat{\mu}_2)^\top \hat{\Sigma}^{-1}(x-\hat{\mu}_2)^\top\right)\hat{\pi}_2}{(2\pi)^{-\frac{p}{2}} \det\left(\hat{\Sigma}^{-\frac{1}{2}}\right) \exp\left(-\frac{1}{2}(x-\hat{\mu}_1)^\top \hat{\Sigma}^{-1}(x-\hat{\mu}_1)^\top\right)\hat{\pi}_1} \\
&= \frac{\exp\left(-\frac{1}{2}(x-\hat{\mu}_2)^\top \hat{\Sigma}^{-1}(x-\hat{\mu}_2)^\top + \log(\hat{\pi}_2)\right)}{\exp\left(-\frac{1}{2}(x-\hat{\mu}_1)^\top \hat{\Sigma}^{-1}(x-\hat{\mu}_1)^\top + \log(\hat{\pi}_1)\right)} \\
&= \exp\left(-\frac{1}{2}x^\top\hat{\Sigma}^{-1}x + x^\top\hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_2^\top\hat{\Sigma}^{-1}\hat{\mu}_2 + \log(\hat{\pi}_2)\right. \\
&\qquad \left. - \left[-\frac{1}{2}x^\top\hat{\Sigma}^{-1}x + x^\top\hat{\Sigma}^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_1^\top\hat{\Sigma}^{-1}\hat{\mu}_1 + \log(\hat{\pi}_1)\right]\right) \\
&= \exp\left(x^\top \underbrace{\Sigma^{-1}(\hat{\mu}_2-\hat{\mu}_1)}_{-\hat{\beta}} \underbrace{- \frac{1}{2}(\hat{\mu}_2+\hat{\mu}_1)\hat{\Sigma}^{-1}(\hat{\mu}_2-\hat{\mu}_1) + \log\left(\frac{\hat{\pi}_2}{\hat{\pi}_1}\right)}_{\hat{\beta}_0}\right)
\end{aligned}
$$

LDA – Logistic Regression

- So, LDA implies the same model for the conditional probabilies $P(Y|X)$ as logistic regression but uses a different method to estimate the parameters
- Interpreted as discriminative classifier, LDA is thus logistic regression with a different loss function
- LDA and Logistic regression are thus a generative-discriminative pair

dis gen

Generative vs Discriminative

- Which one is more accurate?
  - Generative classifier makes more assumptions (namely about $P(X|Y)$, which is assumed to be Gaussian for LDA)
  - -> Generative classifier has lower variance but higher bias
  - -> Generative classifer tends to work better for small N whereas discriminative classfier tends to work better for large N
- Other differences
  - By also estimating $P(X|Y)$, generative classifiers can deal with missing feature values "automatically"
  - LDA and Logistic Regression perform very similarly on most problems
  - Most succesful, modern methods (random forests, support vector machines, deep neural networks) are discriminative

To Focus in the Book

- Correct interpretation of coefficients obtained from logistic regression
- Extension of logistic regression to more than 2 classes (multinomial logistic regression)
- Application of the methods (Section 4.7)
- Importance of standardization for kNN (p. 183)