CS 480 / 680

1/27/26

Lecture 6

Logistic Regression

HW1 - due tomorrow

HW3 - t.b. posted —"—

↳ due Feb 4

Sol 3 Feb 10 Tue

Q1 Feb 12 Thu

all material to L7 - Jan 28

\ gradient descent

Lecture II: Linear regression and classification. Loss functions

Marina Meilă
mmp@uwaterloo.ca

With Thanks to Pascal Poupart & Gautam Kamath
Cheriton School of Computer Science
University of Waterloo

January 12, 2026

Linear predictors generalities

Loss functions

Least squares linear regression
- Linear regresssion as minimizing L_{LS}
- Linear regresssion as maximizing likelihood
- Linear Discriminant Analysis (LDA)
- QDA (Quadratic Discriminant Analysis)
- Logistic Regression
- The PERCEPTRON algorithm

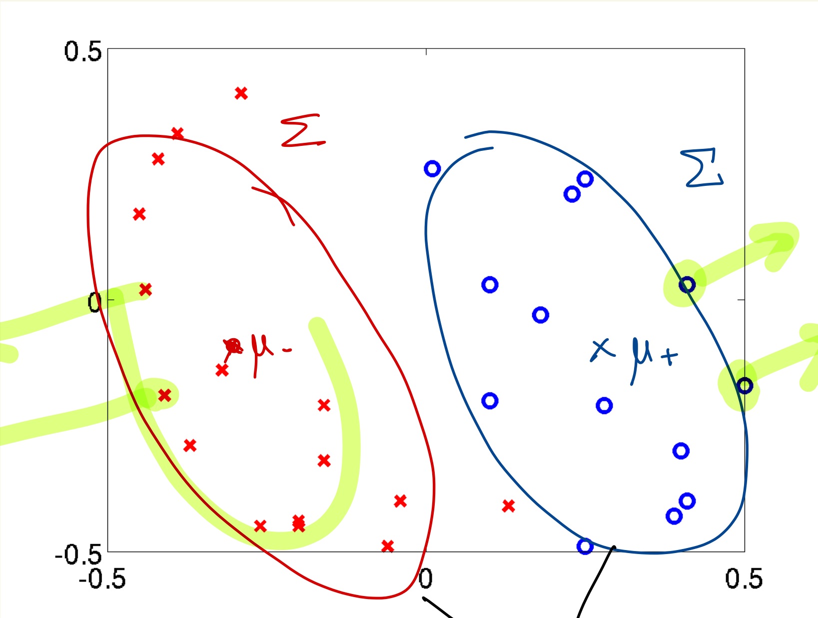
Reading HTF Ch.: 2.1–5,2.9, 7.1–4 bias-variance tradeoff, Murphy Ch.: 1., 8.6[1], Bach Ch.:

[1] Neither textbook is close to these notes except in a few places; take them as alternative perspectives or related reading

Linear classifier

- L_{01} = misclassification error $\longrightarrow$ Perceptron ✓
- Statistical loss "L_{ML}" Max Likelihood $\longrightarrow$ Linear Discriminant Analysis (LDA)
- — " — L_{logit} $\longrightarrow$ Logistic Regression

LDA



Covariance matrix Σ

for calculations

Class + $\rightarrow$ μ_+,

Class − $\rightarrow$ μ_-

$\Big\} \; \Sigma \; (= \sigma^2 I)$

$$\pi_+ = \Pr[y = +] = \frac{n_+}{n}$$

$$n_+ = \#\{ y^i = +1 \}$$

$$\pi_- = 1 - \pi_+$$

$\Pr[y \mid x] = ?$ Bayes' Rule

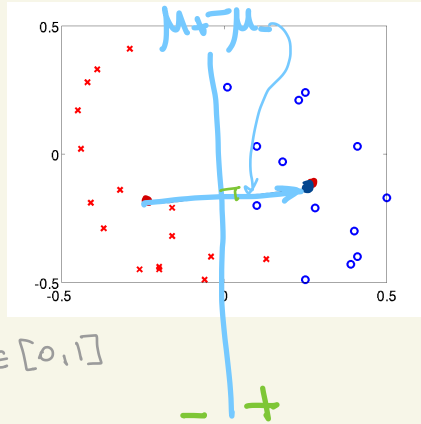
new

Generative Model for classification (NOT Gen Model for unsupervised learning)

1. $y^i \sim (\pi_+, \pi_-) \in \{\pm 1\} \leftarrow P_Y$

2. if $y^i = 1 \Rightarrow x^i \sim N(\mu_+, \Sigma)$
   else $\Rightarrow x^i \sim N(\mu_-, \Sigma)$ $\Big\} \leftarrow P_{X|Y}$

TRAINING: estimated from data



PREDICTION

Given x $\quad P[y|x] = \dfrac{P[x|y]\cdot P_y}{P_{y=+}\cdot P[x|+] + P_{y=-}\cdot P[x|-]}$

wanted

$$f(x) = P[y=+1|x] = \frac{N(\mu_+, \sigma^2 I)\cdot \pi_+}{N(\mu_+, \sigma^2 I)\cdot \pi_+ + N(\mu_-, \sigma^2 I)\pi_-} \in [0,1]$$

$$\hat{y} = \begin{cases} + & \text{if } f(x) \geq 1/2 \\ - & \text{otherwise} \end{cases}$$

STATISTICS $\rightleftarrows$ CALCULUS

$$\pi_+ N(\mu_+, \sigma^2 I) \geq \pi_- N(\mu_-, \sigma^2 I) \quad \Big|\ \ln$$

$$\pi_+ e^{-\frac{\|x-\mu_+\|^2}{2\sigma^2}} \geq \pi_- e^{-\frac{\|x-\mu_-\|^2}{2\sigma^2}}$$

$$\ln \pi_+ - \frac{\|x-\mu_+\|^2}{2\sigma^2} \geq \ln \pi_- - \frac{\|x-\mu_-\|^2}{2\sigma^2}$$

$$\|x-\mu\|^2 = x^Tx + \mu^T\mu - 2\mu^Tx \quad \leftarrow \text{Ex}$$

$$-\|\mu_+\|^2 + \|\mu_-\|^2 + 2\mu_+^Tx - 2\mu_-^Tx \geq \ln\frac{\pi_-}{\pi_+}\cdot 2\sigma^2$$

new x

$$\underbrace{(\mu_+ - \mu_-)^T}_{\beta^T} x \geq \underbrace{-\sigma^2 \ln\frac{\pi_+}{\pi_-} + \frac{\|\mu_+\|^2 - \|\mu_-\|^2}{2}}_{\beta_0} \quad \to \hat{y} = +$$

The PERCEPTRON algorithm

Fitting a linear predictor for classification, first third approach.
Define $f(x) = \beta^T x$ and find β that classifies all the data correctly (when possible).
PERCEPTRON **Algorithm**

Input labeled training set $\mathcal{D}$

Initialize $\beta = 0$, for all i, $x^i \to \frac{x^i}{||x^i||}$ (normalize the inputs)

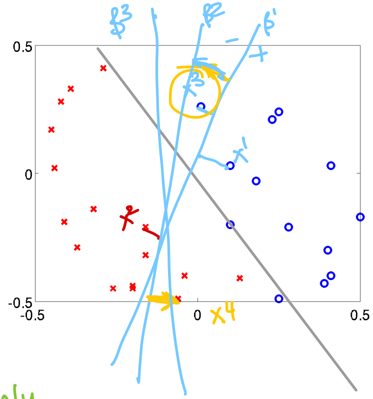
Repeat until no more mistakes

for $i = 1 : N$

1. if $\mathrm{sgn}(\beta^T x^i) \neq y^i$ (a mistake)

$\beta \leftarrow \beta + y^i x^i$

(Other variants exist)



Linearly separable $\mathcal{D}$

The perceptron algorithm and linearly separable data

- $\mathcal{D}$ is **linearly separable** iff there is a β_* so that $\mathrm{sgn} \beta_*^T x^i = y^i$ for all $i = 1, \dots N$.
  If one such β_* exists, then there are an infinity of them

Theorem

Let $\mathcal{D}$ be a linearly separable data set, and define

$$\gamma \;=\; \min_i \frac{|\beta_*^T x^i|}{||\beta_*||||x^i||}. \tag{39}$$

Then, the number of mistakes made by the PERCEPTRON *algorithm is at most $1/\gamma^2$.*

- Note that if we scale the examples to have norm 1, then γ is the minimum distance to the hyperplane $\beta_*^T x = 0$ in the data set.
  Exercise Show that if $\mathcal{D}$ is linearly separable, the scaling $x^i \rightarrow \frac{x^i}{||x^i||}$ leaves it linearly separable.
- If $\mathcal{D}$ is not linearly separable, the algorithm oscillates indefinitely.

Linear Discriminant Analysis (LDA)

Fitting a linear predictor for classification, first approach. (We are in the binary classification case)

- Assume each class is generated by a Normal distribution

$$P_{X|Y}(x|+) = \mathcal{N}(x; \mu_+, \Sigma_+), \quad P_{X|Y}(x|-) = \mathcal{N}(x; \mu_-, \Sigma_-) \quad \text{and } P_Y(1) = p$$

- Given x, what is the probability it came from class $+$?

$$P_{Y|X}(+|x) \;=\; \frac{P_Y(1)P_{X|Y}(x|+)}{P_Y(1)P_{X|Y}(x|+) + P_Y(-)P_{X|Y}(x|+-)} \quad \text{and} \quad P_{Y|X}(-|x) \;=\; 1 - P_{Y|X}(+|x) \quad (19)$$

  This formula is true whether the distributions $P_{X|Y}$ are normal or not.
- We assign x to the class with maximum posterior probability.

$$\hat{y}(x) \;=\; \operatorname*{argmax}_{y\in\{\pm 1} P_{Y|X}(y|x) \quad (20)$$

  This too, holds true whether the distributions $P_{X|Y}$ are normal or not.

LDA – continued

Now we specialize to the case of normal class distribution. We assume in addition that $\Sigma_+ = \Sigma_- = K^{-1}$.

- Decision rule: $\hat{y} = 1$ iff $P_{Y|X}(+|x) > P_{Y|X}(-|x)$
- or equivalently iff

$$
\begin{aligned}
0 \;\leq\; f(x) &= \ln \frac{P_{Y|X}(+|x)}{P_{Y|X}(-|x)} && (21)\\
&= \ln \frac{p}{1-p} - \frac{1}{2}\left[x^T K x - \underline{2\mu_+^T K x} + \mu_+^T K \mu_+\right] \\
&\quad -\frac{1}{2}\left[\cancel{x^T K x} - \underline{2\mu_-^T K x} + \mu_-^T K \mu_-\right] && (22)\\
&= \left[K(\mu_+ - \mu_-)\right]^T x + \ln \frac{p}{1-p} + \frac{\mu_-^T K \mu_- - \mu_+^T K \mu_+}{2} && (23)\\
&= \beta^T x + \beta_0 && (24)
\end{aligned}
$$

- The above is a linear expression in x, hence this classifier is called **(Fisher's) Linear Discriminant**
- Note that if we change the variables to $x \leftarrow \sqrt{K}x$, $\mu_\pm \leftarrow \sqrt{K}\mu_\pm$, and if we shift the origin to $\frac{\mu_+ + \mu_-}{2}$ (24) becomes

$$2\mu_+^T x + \ln \frac{p}{1-p} \qquad (25)$$

This has a geometric interpretation

LDA Algorithm

LDA Algorithm

Train

1. Estimate μ_+ from data $\{(x^i, y^i),\ y^i = +1\}$
2. Estimate μ_- from data $\{(x^i, y^i),\ y^i = -1\}$
3. Estimate Σ jointly for both classes, calculate $K = \Sigma^{-1}$. Exercise Derive the formula for this estimate, in the Max Likelihood setting
4. Estimate $p = |\{(x^i, y^i),\ y^i = +1\}|/n$.

Predict Now apply (24) to classify new data x

And QDA (Quadratic Discriminant Analysis) Not Required

- If we do not assume $\Sigma_+ = \Sigma_-$ then (21) is a quadratic function of x Exercise Plot the curve $f(x) = 0$ in (21) for various data sets in two dimensions. What kind of curves do you observe? Can the decision region be bounded?

$$f(x) =$$

$$= \ln \frac{p}{1-p} - \tfrac{1}{2} \ln |\Sigma_+| + \tfrac{1}{2} \ln |\Sigma_-| \quad (26)$$

$$- \frac{1}{2} \left[x^T \Sigma_+^{-1} x - 2\mu_+^T \Sigma_+^{-1} x + \mu_+^T \Sigma_+^{-1} \mu_+ \right] \quad (27)$$

$$+ \frac{1}{2} \left[x^T \Sigma_-^{-1} x - 2\mu_-^T \Sigma_-^{-1} x + \mu_-^T \Sigma_-^{-1} \mu_- \right] \quad (28)$$

$$= \left[\ln \frac{p}{1-p} - \tfrac{1}{2} \ln |\Sigma_+| + \tfrac{1}{2} \ln |\Sigma_-| - \tfrac{1}{2} \mu_+^T \Sigma_+^{-1} \mu_+ + \tfrac{1}{2} \mu_-^T \Sigma_-^{-1} \mu_- \right] \quad (29)$$

$$+ \underbrace{\left[\mu_+^T \Sigma_+^{-1} - \mu_-^T \Sigma_-^{-1} \right] x}_{\text{linear}} - \underbrace{\frac{1}{2} x^T \left[\Sigma_+^{-1} - \Sigma_-^{-1} \right] x}_{\text{quadratic}} \quad (30)$$

Logistic Regression

$$\left[y_* = \begin{cases} 1 & \text{if } y=1 \\ 0 & \text{if } y=-1 \end{cases} \right]$$

Notation

Fitting a linear predictor for classification, another approach.
Let $\boxed{f(x) = \beta^T x}$ model the **log odds** of class 1

$$f(X) = \ln \frac{\boxed{P(Y = 1|X)}}{P(Y = -1|X)} \tag{31}$$

p

Then

- $\hat{y} = 1$ iff $P(Y = 1|X) > P(Y = -1|X)$
  - just like in the previous case! so what's the difference?

Almost Likelihood

Step 1

for any $x \to f, p$

$$f = \ln \frac{p}{1-p} \to e^f = \frac{p}{1-p} \to (1-p)e^f = p$$

$$\Rightarrow p(1+e^f) = e^f$$

Logit

$$\Rightarrow p = \frac{e^f}{1+e^f} = P[y=1|x]$$

$$1-p = \frac{1}{1+e^f} = P[y\neq 1|x]$$

$$= \frac{e^{f/2}}{e^{-f/2}+e^{f/2}}$$

$$\left[P[y|x] = \frac{e^{y_* f}}{1+e^f} \right]$$

Logistic Regression

Fitting a linear predictor for classification, another approach.
Let $f(x) = \beta^T x$ model the **log odds** of class 1

$$f(X) \;=\; \frac{P(Y=1|X)}{P(Y=-1|X)} \tag{31}$$

Then

- $\hat{y} = 1$ iff $P(Y=1|X) > P(Y=-1|X)$
  - just like in the previous case! so what's the difference?
  - Answer: We don't assume each class is Gaussian, so we are in a more general situation than LDA
- What is $p(x) = P(Y=1|X=x)$ under our linear model?

$$\ln\frac{p}{1-p} \;=\; f, \quad \frac{p}{1-p} \;=\; e^f, \quad p \;=\; \frac{e^f}{1+e^f} \quad 1-p \;=\; \frac{1}{1+e^f} \tag{32}$$

An alternative "symmetric" expression for $p, 1-p$ is

$$p \;=\; \frac{e^{f/2}}{e^{f/2}+e^{-f/2}}, \quad 1-p \;=\; \frac{e^{-f/2}}{e^{f/2}+e^{-f/2}}. \tag{33}$$

Estimating the parameters by Max Likelihood

Step 2. Log-likelihood ℓ

- Denote $y_* = (1-y)/2 \in \{0,1\}$
- The likelihood of a data point is $\boxed{P_{Y|X}(y|x) = \frac{e^{y_* f(x)}}{1+e^{f(x)}}}$ $= L$ for $n=1$
- The log-likelihood is $l(\beta; x) = y_* f(x) - \ln(1+e^{f(x)})$
- $\frac{\partial l}{\partial f} = y_* - \frac{e^f}{1+e^f} = y_* - \frac{1}{1+e^{-f}}$
  This is a scalar, and $\operatorname{sgn} \frac{\partial l}{\partial f} = y$
- We have also $\frac{\partial f(x)}{\partial \beta} = x$
- Now, the gradient of l w.r.t the parameter vector β is

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial f}\frac{\partial f}{\partial \beta} = \left(y_* - \frac{1}{1+e^{-f(x)}}\right)x \tag{34}$$

Interpretation: The infinitezimal change of β to increase log-likelihood for a single data point is along the direction of x, with the sign of y

Likelihood $L(\beta) = P[y|x,\beta] = P[y|x]$

$\mathcal{D} = \{(x^i, y^i),\ i=1:n\}$

Log-likelihood $\ell(\beta) = \sum_{i=1}^{n}\left[y_*^i f(x^i) - \ln(1+e^{f(x^i)})\right]$

START

GRAD ASCENT

Step 3. Gradient

Gradient of ℓ w.r.t β

$n=1$

$$\ell = \underline{y_* f(x)} - \underbrace{\ln\left(1+e^{f(x)}\right)}$$

$$\frac{\partial \ell}{\partial f} = y_* - \frac{e^f}{\underbrace{1+e^f}_{\mathbb{P}[y=+]}} \quad \in \mathbb{R}$$

$$\frac{\partial f}{\partial \beta} = \nabla f = \nabla_\beta \left(x^T \beta\right) = x \quad \in \mathbb{R}^d$$

$$\nabla \ell \equiv \frac{\partial \ell}{\partial \beta} = \left(y_* - \frac{e^f}{1+e^f}\right) x = y \cdot w \cdot x$$

± 1

$y=+ \quad y_* = 1 \Rightarrow \overbrace{1 - P[y^i = 1 \mid x^i]}^{w^i}$

$y=- \quad y_* = 0 \Rightarrow -P[y=1] = -\left(\underbrace{1 - P[y_* = 0]}_{w^i}\right)$

Step 4. (next lecture) **Gradient ascent**

$$\beta \leftarrow \beta + \eta \frac{\partial \ell}{\partial \beta}$$

$\eta > 0$ "step size"

Estimating the parameters by Max Likelihood

- Denote $y_* = (1-y)/2 \in \{0,1\}$
- The likelihood of a data point is $P_{Y|X}(y|x) = \frac{e^{y_* f(x)}}{1+e^{f(x)}}$
- The log-likelihood is $l(\beta; x) = y_* f(x) - \ln(1+e^{f(x)})$
- $\frac{\partial l}{\partial f} = y_* - \frac{e^f}{1+e^f} = y_* - \frac{1}{1+e^{-f}}$
  This is a scalar, and $\operatorname{sgn} \frac{\partial l}{\partial f} = y$
- We have also $\frac{\partial f(x)}{\partial \beta} = x$
- Now, the gradient of l w.r.t the parameter vector β is

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial f}\frac{\partial f}{\partial \beta} = \left(y_* - \frac{1}{1+e^{-f(x)}} \right) x \tag{34}$$

  Interpretation: The infinitezimal change of β to increase log-likelihood for a single data point is along the direction of x, with the sign of y
- Log-likelihood of the data set $\mathcal{D}$

$$l(\beta; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{d} l(\beta; (x^i, y^i)) \tag{35}$$

- The optimal β maximizes $l(\beta; \mathcal{D})$ and therefore

$$\frac{\partial l(\beta; \mathcal{D})}{\partial \beta} = \frac{1}{N} \sum_{i=1}^{d} \left(y_*^i - \frac{1}{1+e^{-f(x^i)}} \right) x^i = 0 \tag{36}$$

- Unfortunately, (36) does not have a closed form solution!
  We maximize the (log)likelihood by iterative methods (e.g. gradient ascent) to obtain the β of the classifier.

The gradient – an alternative formula

- We use the original y values instead of y_*
- Note that

$$P_{Y|X}(y|x) = \frac{1}{1+e^{-yf(x)}} = \phi(yf(x)) \quad (37)$$

- with $\phi' = \phi(1-\phi)$
- Then, $\frac{\partial \ln P_{Y|X}(y|x)}{\partial f} = \frac{\partial \ln \phi(yf)}{\partial f} = \frac{y\phi(yf)(1-\phi(yf))}{\phi(yf)} = y(1-\phi(yf))$
- The gradient of the log-likelihood of the data is now

$$\frac{\partial l(\beta;\mathcal{D})}{\partial \beta} = \frac{1}{N}\sum_{i=1}^{d}\left(1-\underbrace{\phi(e^{yf(x^i)})}_{P_{Y|X}(y_i|x^i,\beta)}\right)y_i x^i \quad (38)$$