Linear Models for Regression and Classification, Learning

Tomáš Svoboda and Petr Pošík
thanks to Matěj Hoffmann, Daniel Novák, Filip Železný, Ondřej Drbohlav

Vision for Robots and Autonomous Systems, Center for Machine Perception
Department of Cybernetics
Faculty of Electrical Engineering, Czech Technical University in Prague

May 18, 2023

Supervised Learning

A training multi-set of examples is available. Correct answers (hidden state, class, the quantity we want to predict) are *known* for all training examples.

Classification :

- Nominal dependent variable
- Examples: predict spam/ham based on email contents, predict 0/1/.../9 based on the image of a number, etc.

Regression :

- Quantitative/continuous dependent variable
- Examples: predict temperature in Prague based on date and time, predict height of a person based on weight and gender, etc.

Supervised Learning

A training multi-set of examples is available. Correct answers (hidden state, class, the quantity we want to predict) are *known* for all training examples.

Classification :

- Nominal dependent variable
- Examples: predict spam/ham based on email contents, predict $0/1/\ldots/9$ based on the image of a number, etc.

Regression :

- Quantitative/continuous dependent variable
- Examples: predict temperature in Prague based on date and time, predict height of a person based on weight and gender, etc.

Supervised Learning

A training multi-set of examples is available. Correct answers (hidden state, class, the quantity we want to predict) are *known* for all training examples.

Classification :

- Nominal dependent variable
- Examples: predict spam/ham based on email contents, predict $0/1/\ldots/9$ based on the image of a number, etc.

Regression :

- Quantitative/continuous dependent variable
- Examples: predict temperature in Prague based on date and time, predict height of a person based on weight and gender, etc.

Learning by minimization of empirical risk

- Given the set of parametrized strategies $\delta\colon \mathcal{X} \to \mathcal{D}$, penalty/loss function $\ell\colon \mathcal{S} \times \mathcal{D} \to \mathbb{R}$, the quality of each strategy δ could be described by the risk

$$R(\delta) = \sum_{s \in \mathcal{S}} \sum_{x \in \mathcal{X}} P(x, s)\ell(s, \delta(x)),$$

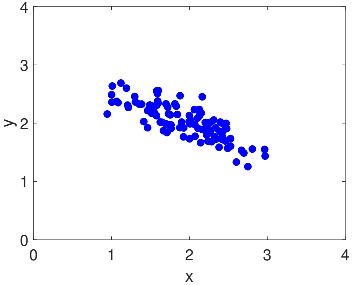
  but P is unknown.
- We thus use the empirical risk R_{emp} error on training (multi)set $\mathcal{T} = \{(x^{(i)}, s^{(i)})\}_{i=1}^{N}$, $x \in \mathcal{X}$, $s \in \mathcal{S}$:

$$R_{\mathrm{emp}}(\delta) = \frac{1}{N} \sum_{(x^{(i)}, s^{(i)}) \in \mathcal{T}} \ell(s^{(i)}, \delta(x^{(i)})).$$

- Optimal strategy $\delta^* = \operatorname{argmin}_{\delta} R_{\mathrm{emp}}(\delta)$.
- We expect the data are from the right distribution.

Quiz: Line fitting

We would like to fit a line of the form $\widehat{y} = w_0 + w_1 x$ to the following data:



The parameters of a line with a good fit will likely be

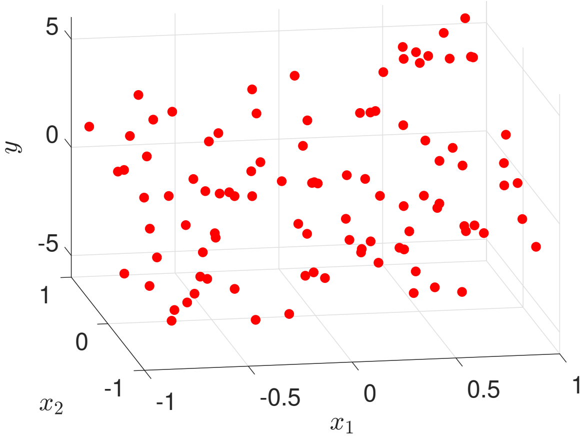
A $w_0 = -1,\ w_1 = -2$

B $w_0 = -\frac{1}{2},\ w_1 = 1$

C $w_0 = 3,\ w_1 = -\frac{1}{2}$

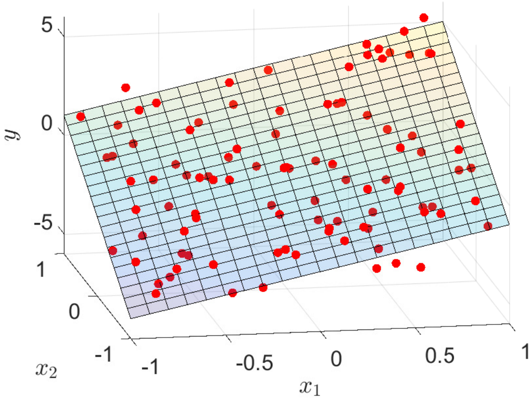
D $w_0 = 2,\ w_1 = \frac{1}{3}$

Linear regression: Illustration



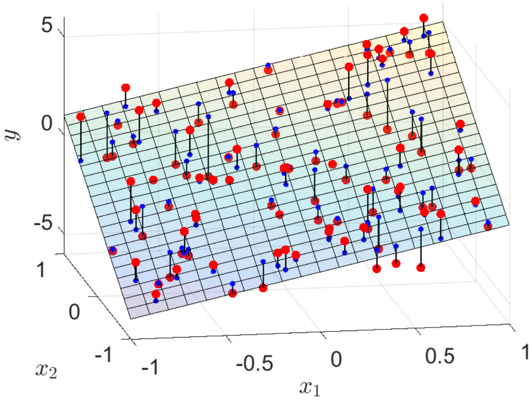
Given a dataset of input vectors $\boldsymbol{x}^{(i)}$ and the respective values of output variable $y^{(i)}$ ...

Linear regression: Illustration



...we would like to find a linear model of this dataset ...

Linear regression: Illustration



...minimizing the errors between target values and the model predictions.

Regression

Reformulating Linear algebra in a machine learning language.

Regression task is a supervised learning task, i.e.

- a training (multi)set $\mathcal{T} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \dots, (\boldsymbol{x}^{(N)}, y^{(N)})\}$ is available, where
- the labels $y^{(i)}$ are *quantitative*, often *continuous* (as opposed to classification tasks where $y^{(i)}$ are nominal).
- Its purpose is to model the relationship between independent variables (inputs) $\boldsymbol{x} = (x_1, \dots, x_D)$ and the dependent variable (output) y.

Linear Regression

Linear regression is a particular regression model which assumes (and learns) linear relationship between the inputs and the output:

$$\widehat{y} = \delta(\boldsymbol{x}) = w_0 + w_1 x_1 + \ldots + w_D x_D = w_0 + \langle \boldsymbol{w}, \boldsymbol{x} \rangle = w_0 + \boldsymbol{w}^\top \boldsymbol{x},$$

where

- $\widehat{y}$ is the model *prediction* (*estimate* of the true value y),
- $\delta(\boldsymbol{x})$ is the decision strategy (a linear model in this case),
- $w_0, \ldots, w_D$ are the coefficients of the linear function (weights), w_0 is the *bias*,
- $\langle \boldsymbol{w}, \boldsymbol{x} \rangle$ is a *dot product* of vectors $\boldsymbol{w}$ and $\boldsymbol{x}$ (scalar product),
- which can be also computed as a matrix product $\boldsymbol{w}^\top \boldsymbol{x}$ if $\boldsymbol{w}$ and $\boldsymbol{x}$ are *column vectors*, i.e. matrices of size $[D \times 1]$.

Notation remarks

Homogeneous coordinates :

- If we add "1" as the first element of $\boldsymbol{x}$ so that $\boldsymbol{x} = (1, x_1, \ldots, x_D)$, and
- include the bias term w_0 in the vector $\boldsymbol{w}$ so that $\boldsymbol{w} = (w_0, w_1, \ldots, w_D)$, then

$$\widehat{y} = \delta(\boldsymbol{x}) = w_0 \cdot 1 + w_1 x_1 + \ldots + w_D x_D = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \boldsymbol{w}^\top \boldsymbol{x}.$$

Matrix notation: If we organize the data $\mathcal{T}$ into matrices $\boldsymbol{X}$ and $\boldsymbol{y}$, such that

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ldots & 1 \\ \boldsymbol{x}^{(1)} & \ldots & \boldsymbol{x}^{(N)} \end{pmatrix} \quad \text{and} \quad \boldsymbol{y} = \left(y^{(1)}, \ldots, y^{(N)}\right),$$

and similarly with $\widehat{\boldsymbol{y}}$, then we can write a batch computation of predictions for all data in $\boldsymbol{X}$ as

$$\widehat{\boldsymbol{y}} = \left(\delta(\boldsymbol{x}^{(1)}), \ldots, \delta(\boldsymbol{x}^{(N)})\right) = \left(\boldsymbol{w}^\top \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{w}^\top \boldsymbol{x}^{(N)}\right) = \boldsymbol{w}^\top \boldsymbol{X}.$$

Notation remarks

Homogeneous coordinates :

- If we add "1" as the first element of $\boldsymbol{x}$ so that $\boldsymbol{x} = (1, x_1, \ldots, x_D)$, and
- include the bias term w_0 in the vector $\boldsymbol{w}$ so that $\boldsymbol{w} = (w_0, w_1, \ldots, w_D)$, then

$$\widehat{y} = \delta(\boldsymbol{x}) = w_0 \cdot 1 + w_1 x_1 + \ldots + w_D x_D = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \boldsymbol{w}^\top \boldsymbol{x}.$$

Matrix notation: If we organize the data $\mathcal{T}$ into matrices $\boldsymbol{X}$ and $\boldsymbol{y}$, such that

$$\boldsymbol{X} = \begin{pmatrix} 1 & \ldots & 1 \\ \boldsymbol{x}^{(1)} & \ldots & \boldsymbol{x}^{(N)} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{y} = \left(y^{(1)}, \ldots, y^{(N)} \right),$$

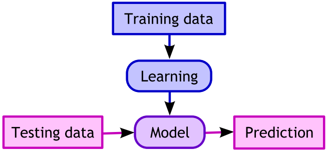
and similarly with $\widehat{\boldsymbol{y}}$, then we can write a batch computation of predictions for all data in $\boldsymbol{X}$ as

$$\widehat{\boldsymbol{y}} = \left(\delta(\boldsymbol{x}^{(1)}), \ldots, \delta(\boldsymbol{x}^{(N)}) \right) = \left(\boldsymbol{w}^\top \boldsymbol{x}^{(1)}, \ldots, \boldsymbol{w}^\top \boldsymbol{x}^{(N)} \right) = \boldsymbol{w}^\top \boldsymbol{X}.$$

Two operation modes

Any ML model has 2 operation modes:

1. learning (training, fitting) of δ and
2. application of δ (testing, making predictions).



The dec. strategy δ can be viewed as a function of 2 variables: $\delta(\boldsymbol{x}, \boldsymbol{w})$.

Model application: (Inference) Given $\boldsymbol{w}$, we can manipulate $\boldsymbol{x}$ to make predictions:

$$\widehat{y} = \delta(\boldsymbol{x}, \boldsymbol{w}) = \delta_{\boldsymbol{w}}(\boldsymbol{x}).$$

Model learning: Given $\mathcal{T}$, we can tune the model parameters $\boldsymbol{w}$ to fit the model to the data:

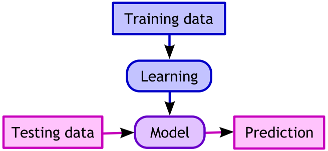
$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}} R_{\mathrm{emp}}(\delta_{\boldsymbol{w}}) = \operatorname*{argmin}_{\boldsymbol{w}} J(\boldsymbol{w}, \mathcal{T})$$

$J(\boldsymbol{w}, \mathcal{T})$ and $\ell(\boldsymbol{w}, \mathcal{T})$ are closely related. Optimization criterium $J()$ is a broader term. $\ell()$ essentially measures discrepancy between true data and the predictions. How to train the model?

Two operation modes

Any ML model has 2 operation modes:

1. learning (training, fitting) of δ and
2. application of δ (testing, making predictions).



The dec. strategy δ can be viewed as a function of 2 variables: $\delta(\boldsymbol{x}, \boldsymbol{w})$.

Model application: (Inference) Given $\boldsymbol{w}$, we can manipulate $\boldsymbol{x}$ to make predictions:

$$\widehat{y} = \delta(\boldsymbol{x}, \boldsymbol{w}) = \delta_{\boldsymbol{w}}(\boldsymbol{x}).$$

Model learning: Given $\mathcal{T}$, we can tune the model parameters $\boldsymbol{w}$ to fit the model to the data:

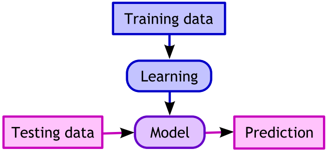
$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}} R_{\mathrm{emp}}(\delta_{\boldsymbol{w}}) = \operatorname*{argmin}_{\boldsymbol{w}} J(\boldsymbol{w}, \mathcal{T})$$

$J(\boldsymbol{w}, \mathcal{T})$ and $\ell(\boldsymbol{w}, \mathcal{T})$ are closely related. Optimization criterium $J()$ is a broader term. $\ell()$ essentially measures discrepancy between true data and the predictions. How to train the model?

Two operation modes

Any ML model has 2 operation modes:

1. learning (training, fitting) of δ and
2. application of δ (testing, making predictions).



The dec. strategy δ can be viewed as a function of 2 variables: $\delta(\boldsymbol{x}, \boldsymbol{w})$.

Model application: (Inference) Given $\boldsymbol{w}$, we can manipulate $\boldsymbol{x}$ to make predictions:

$$\widehat{y} = \delta(\boldsymbol{x}, \boldsymbol{w}) = \delta_{\boldsymbol{w}}(\boldsymbol{x}).$$

Model learning: Given $\mathcal{T}$, we can tune the model parameters $\boldsymbol{w}$ to fit the model to the data:

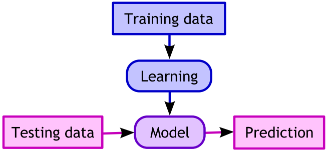
$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, R_{\mathrm{emp}}(\delta_{\boldsymbol{w}}) = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, J(\boldsymbol{w}, \mathcal{T})$$

$J(\boldsymbol{w}, \mathcal{T})$ and $\ell(\boldsymbol{w}, \mathcal{T})$ are closely related. Optimization criterium $J()$ is a broader term. $\ell()$ essentially measures discrepancy between true data and the predictions. How to train the model?

Two operation modes

Any ML model has 2 operation modes:

1. learning (training, fitting) of δ and
2. application of δ (testing, making predictions).



The dec. strategy δ can be viewed as a function of 2 variables: $\delta(\boldsymbol{x}, \boldsymbol{w})$.

Model application: (Inference) Given $\boldsymbol{w}$, we can manipulate $\boldsymbol{x}$ to make predictions:

$$\widehat{y} = \delta(\boldsymbol{x}, \boldsymbol{w}) = \delta_{\boldsymbol{w}}(\boldsymbol{x}).$$

Model learning: Given $\mathcal{T}$, we can tune the model parameters $\boldsymbol{w}$ to fit the model to the data:

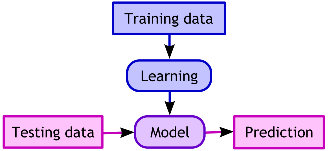
$$\boldsymbol{w}^* = \underset{\boldsymbol{w}}{\text{argmin}}\, R_{\text{emp}}(\delta_{\boldsymbol{w}}) = \underset{\boldsymbol{w}}{\text{argmin}}\, J(\boldsymbol{w}, \mathcal{T})$$

$J(\boldsymbol{w}, \mathcal{T})$ and $\ell(\boldsymbol{w}, \mathcal{T})$ are closely related. Optimization criterium $J()$ is a broader term. $\ell()$ essentially measures discrepancy between true data and the predictions. How to train the model?

Two operation modes

Any ML model has 2 operation modes:

1. learning (training, fitting) of δ and
2. application of δ (testing, making predictions).



The dec. strategy δ can be viewed as a function of 2 variables: $\delta(\boldsymbol{x}, \boldsymbol{w})$.

Model application: (Inference) Given $\boldsymbol{w}$, we can manipulate $\boldsymbol{x}$ to make predictions:

$$\widehat{y} = \delta(\boldsymbol{x}, \boldsymbol{w}) = \delta_{\boldsymbol{w}}(\boldsymbol{x}).$$

Model learning: Given $\mathcal{T}$, we can tune the model parameters $\boldsymbol{w}$ to fit the model to the data:

$$\boldsymbol{w}^* = \operatorname*{argmin}_{\boldsymbol{w}} R_{\mathrm{emp}}(\delta_{\boldsymbol{w}}) = \operatorname*{argmin}_{\boldsymbol{w}} J(\boldsymbol{w}, \mathcal{T})$$

$J(\boldsymbol{w}, \mathcal{T})$ and $\ell(\boldsymbol{w}, \mathcal{T})$ are closely related. Optimization criterium $J()$ is a broader term. $\ell()$ essentially measures discrepancy between true data and the predictions. How to train the model?

Simple (univariate) linear regression

Simple regression

- $\boldsymbol{x}^{(i)} = x^{(i)}$, i.e., the examples are described by a single feature (they are 1-dimensional).
- Find parameters w_0, w_1 of a linear model $\hat{y} = w_0 + w_1 x$ given a training (multi)set $\mathcal{T} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$.

How to fit a line depending on the number of training examples N:

- $N = 1$ (1 equation, 2 parameters) $\Rightarrow \infty$ linear functions with zero error
- $N = 2$ (2 equations, 2 parameters) $\Rightarrow$ 1 linear function with zero error
- $N \geq 3$ (> 2 equations, 2 parameters) $\Rightarrow$ no linear function with zero error (in general) $\Rightarrow$ a line which minimizes the "size" of error $y - \widehat{y}$ can be fitted:

$$\boldsymbol{w}^* = (w_0^*, w_1^*) = \operatorname*{argmin}_{w_0, w_1} R_{\mathrm{emp}}(w_0, w_1) = \operatorname*{argmin}_{w_0, w_1} J(w_0, w_1, \mathcal{T}).$$

Simple (univariate) linear regression

Simple regression

- $\boldsymbol{x}^{(i)} = x^{(i)}$, i.e., the examples are described by a single feature (they are 1-dimensional).
- Find parameters w_0, w_1 of a linear model $\hat{y} = w_0 + w_1 x$ given a training (multi)set $\mathcal{T} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$.

How to fit a line depending on the number of training examples N:

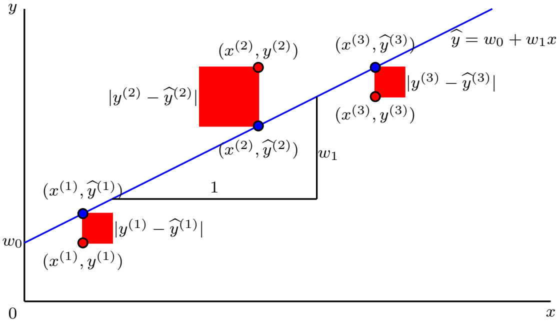
- $N = 1$ (1 equation, 2 parameters) $\Rightarrow \infty$ linear functions with zero error
- $N = 2$ (2 equations, 2 parameters) $\Rightarrow$ 1 linear function with zero error
- $N \geq 3$ (> 2 equations, 2 parameters) $\Rightarrow$ no linear function with zero error (in general) $\Rightarrow$ a line which minimizes the "size" of error $y - \widehat{y}$ can be fitted:

$$\boldsymbol{w}^* = (w_0^*, w_1^*) = \underset{w_0, w_1}{\operatorname{argmin}} R_{\text{emp}}(w_0, w_1) = \underset{w_0, w_1}{\operatorname{argmin}} J(w_0, w_1, \mathcal{T}).$$

The least squares method

Choose such parameters $\boldsymbol{w}$ which minimize the *mean squared error* (MSE)

$$J_{MSE}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \widehat{y}^{(i)}\right)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \delta_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})\right)^2 .$$



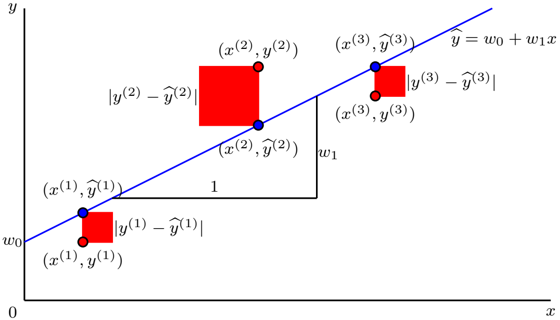
Is there a (closed-form) solution? Explicit solution:

$$w_1 = \frac{\sum_{i=1}^{N}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^{N}(x^{(i)} - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{\text{covariance of } X \text{ and } Y}{\text{variance of } X} \qquad w_0 = \bar{y} - w_1\bar{x}$$

The least squares method

Choose such parameters $\boldsymbol{w}$ which minimize the *mean squared error* (MSE)

$$J_{MSE}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \widehat{y}^{(i)}\right)^2$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \delta_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})\right)^2 .$$



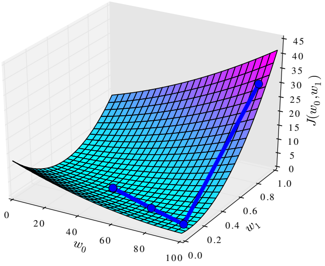
Is there a (closed-form) solution? Explicit solution:

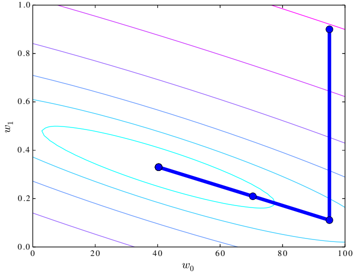
$$w_1 = \frac{\sum_{i=1}^{N}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum_{i=1}^{N}(x^{(i)} - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{\text{covariance of } X \text{ and } Y}{\text{variance of } X}$$

$$w_0 = \bar{y} - w_1\bar{x}$$

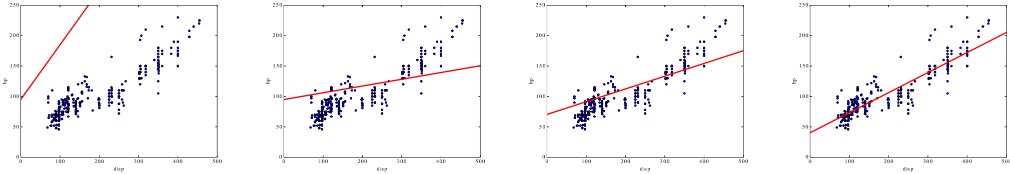
Universal fitting method: minimization of cost function J

The landscape of J in the space of parameters w_0 and w_1:





Gradually better linear models found by an optimization method (BFGS):



Gradient descent algorithm

Given a function $J(w_0, w_1)$ that should be minimized,

- start with a guess of w_0 and w_1 and
- change it, so that $J(w_0, w_1)$ decreases, i.e.
- update our current guess of w_0 and w_1 by taking a step in the direction opposite to the gradient:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \nabla J(w_0, w_1), \text{ i.e.}$$
$$w_i \leftarrow w_i - \alpha \frac{\partial}{\partial w_i} J(w_0, w_1),$$

where all w_is are updated simultaneously and α is a learning rate (step size).

Gradient descent for MSE minimization

For the cost function

$$J(w_0, w_1) = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - \delta_{\boldsymbol{w}}(x^{(i)}) \right)^2 = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - (w_0 + w_1 x^{(i)}) \right)^2 ,$$

the gradient can be computed as

$$\frac{\partial}{\partial w_0} J(w_0, w_1) = -\frac{2}{N} \sum_{i=1}^{N} \left(y^{(i)} - \delta_{\boldsymbol{w}}(x^{(i)}) \right)$$

$$\frac{\partial}{\partial w_1} J(w_0, w_1) = -\frac{2}{N} \sum_{i=1}^{N} \left(y^{(i)} - \delta_{\boldsymbol{w}}(x^{(i)}) \right) x^{(i)}$$

Multivariate linear regression

- $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \ldots, x_D^{(i)})^\top$, i.e. the examples are described by more than 1 feature (they are D-dimensional).
- Find parameters $\boldsymbol{w} = (w_0, \ldots, w_D)^\top$ of a linear model $\widehat{y} = \boldsymbol{w}^\top \boldsymbol{x}$ given the training (multi)set $\mathcal{T} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^N$.

Training: foreach (i): $y^{(i)} = \boldsymbol{w}^\top \boldsymbol{x}^{(i)}$.
In the matrix form:

$$\boldsymbol{y} = \boldsymbol{w}^\top \boldsymbol{X}$$

What is the dimension of $\boldsymbol{X}$?

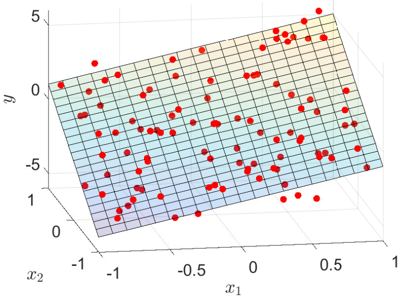
A $(D+1) \times (D+1)$

B $(D+1) \times N$

C $N \times (D+1)$

D $N \times N$

The model is a *hyperplane* in the $(D+1)$ dimensional space.



Multivariate linear regression

- $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \dots, x_D^{(i)})^\top$, i.e. the examples are described by more than 1 feature (they are D-dimensional).
- Find parameters $\boldsymbol{w} = (w_0, \dots, w_D)^\top$ of a linear model $\widehat{y} = \boldsymbol{w}^\top \boldsymbol{x}$ given the training (multi)set $\mathcal{T} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$.

Training: foreach (i): $y^{(i)} = \boldsymbol{w}^\top \boldsymbol{x}^{(i)}$.
In the matrix form:

$$\boldsymbol{y} = \boldsymbol{w}^\top \boldsymbol{X}$$

What is the dimension of $\boldsymbol{X}$?

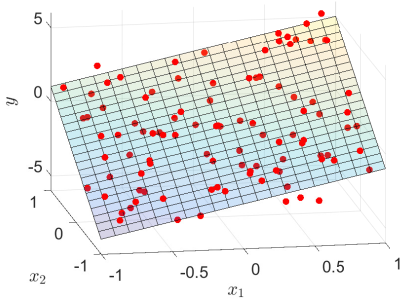
A $(D+1) \times (D+1)$

B $(D+1) \times N$

C $N \times (D+1)$

D $N \times N$

The model is a *hyperplane* in the $(D+1)$ dimensional space.



Multivariate linear regression: learning

1. Numeric optimization of $J(\boldsymbol{w}, T)$:
   - Works as for simple regression, it only searches a space with more dimensions.
   - Sometimes one needs to tune some parameters of the optimization algorithm to work properly (learning rate in gradient descent, etc.).
   - May be slow (many iterations needed), but works even for very large D.

2. Normal equation:

$$\boldsymbol{w}^* = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{y}^\top$$

   - Method to solve for the optimal $\boldsymbol{w}^*$ analytically!
   - No need to choose optimization algorithm parameters. No iterations.
   - Needs to compute $(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}$, which is $O((D+1)^3)$. Becomes intractable for large D.

Multivariate linear regression: learning

1. Numeric optimization of $J(\boldsymbol{w}, T)$:
   - Works as for simple regression, it only searches a space with more dimensions.
   - Sometimes one needs to tune some parameters of the optimization algorithm to work properly (learning rate in gradient descent, etc.).
   - May be slow (many iterations needed), but works even for very large D.

2. Normal equation:

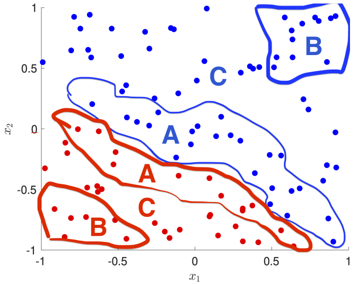
$$\boldsymbol{w}^* = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{y}^\top$$

   - Method to solve for the optimal $\boldsymbol{w}^*$ analytically!
   - No need to choose optimization algorithm parameters. No iterations.
   - Needs to compute $(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}$, which is $O((D+1)^3)$. Becomes intractable for large D.

Classification

- Binary classification
- Discriminant function
- Classification as a regression problem (linear, logistic regression)
- What is the right loss function?
- Etalon classifier (meeting nearest neighbour and linear classifier)
- Acuracy vs precision

Quiz: Importance of training examples



Intuitively, which of the training data points should have the biggest influence on the decision whether a new, unlabeled data point shall be red or blue?

A Those which are closest to data points with the opposite color.

B Those which are farthest from the data points of the opposite color.

C Those which are near the middle of the points with the same color.

D None. All of the data points have the same importance.

Binary classification task

Let's have a training dataset $T = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})$:

- each example described by a vector $\boldsymbol{x} = (x_1, \ldots, x_D)$,
- labeled with the correct class $y \in \{+1, -1\}$.

The goal:

- Find the classifier (decision strategy/rule) δ that minimizes the empirical risk $R_{\mathrm{emp}}(\delta)$.

Discriminant function

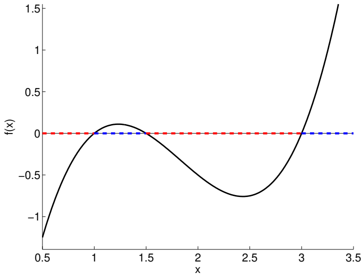
Discriminant function $f(\boldsymbol{x})$:

- It assigns a real number to each observation $\boldsymbol{x}$, may be linear or non-linear.
- For 2 classes, 1 discriminant function is enough.
- It is used to create a decision rule (which then assigns a class to an observation):

$$\widehat{y} = \delta(\boldsymbol{x}) = \begin{cases} +1 & \text{iff} \quad f(\boldsymbol{x}) > 0, \textit{ and} \\ -1 & \text{iff} \quad f(\boldsymbol{x}) < 0. \end{cases}$$

i.e. $\widehat{y} = \delta(\boldsymbol{x}) = \operatorname{sign}(f(\boldsymbol{x}))$.

- Decision boundary: $\{\boldsymbol{x} | f(\boldsymbol{x}) = 0\}$
- Linear classification: the decision boundaries must be linear.
- *Learning* then amounts to finding (suitable parameters of) function f.



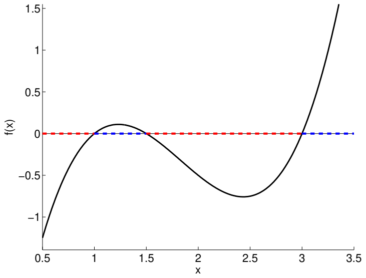
Discriminant function

Discriminant function $f(\boldsymbol{x})$:

- It assigns a real number to each observation $\boldsymbol{x}$, may be linear or non-linear.
- For 2 classes, 1 discriminant function is enough.
- It is used to create a decision rule (which then assigns a class to an observation):

$$\widehat{y} = \delta(\boldsymbol{x}) = \begin{cases} +1 & \text{iff} \quad f(\boldsymbol{x}) > 0, \textit{ and} \\ -1 & \text{iff} \quad f(\boldsymbol{x}) < 0. \end{cases}$$

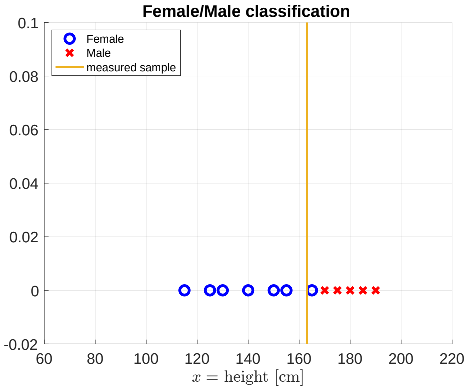
  i.e. $\widehat{y} = \delta(\boldsymbol{x}) = \mathrm{sign}\,(f(\boldsymbol{x}))$.
- Decision boundary: $\{\boldsymbol{x} | f(\boldsymbol{x}) = 0\}$
- Linear classification: the decision boundaries must be linear.
- *Learning* then amounts to finding (suitable parameters of) function f.



Example: Female/Male classification based on height

Training (multi)set $\mathcal{T} = \{(x^{(i)}, s^{(i)})\}_{i=1}^{N}$, $x^{(i)} \in \mathcal{X}$, $s^{(i)} \in \mathcal{S} = \{F, M\}$

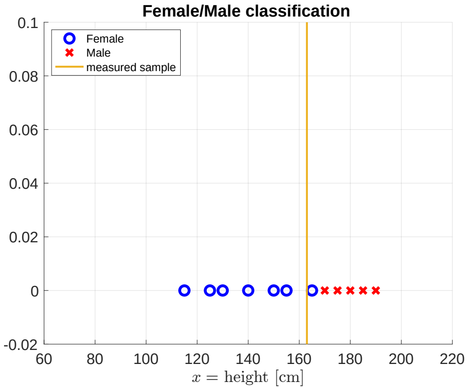
i	1	2	3	4	5	6	7	8	9	10	11	12
Height $x^{(i)}$	115	125	130	140	150	155	165	170	175	180	185	190
Gender $s^{(i)}$	F	F	F	F	F	F	F	M	M	M	M	M



Example: Female/Male classification based on height

Training (multi)set $\mathcal{T} = \{(x^{(i)}, s^{(i)})\}_{i=1}^{N}$, $x^{(i)} \in \mathcal{X}$, $s^{(i)} \in \mathcal{S} = \{F, M\}$

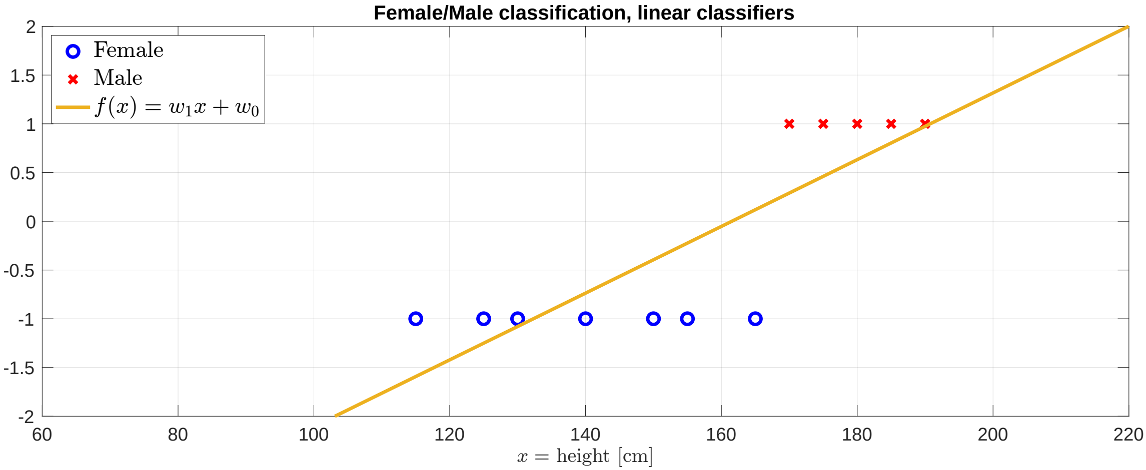
i	1	2	3	4	5	6	7	8	9	10	11	12
Height $x^{(i)}$	115	125	130	140	150	155	165	170	175	180	185	190
Gender $s^{(i)}$	F	F	F	F	F	F	F	M	M	M	M	M



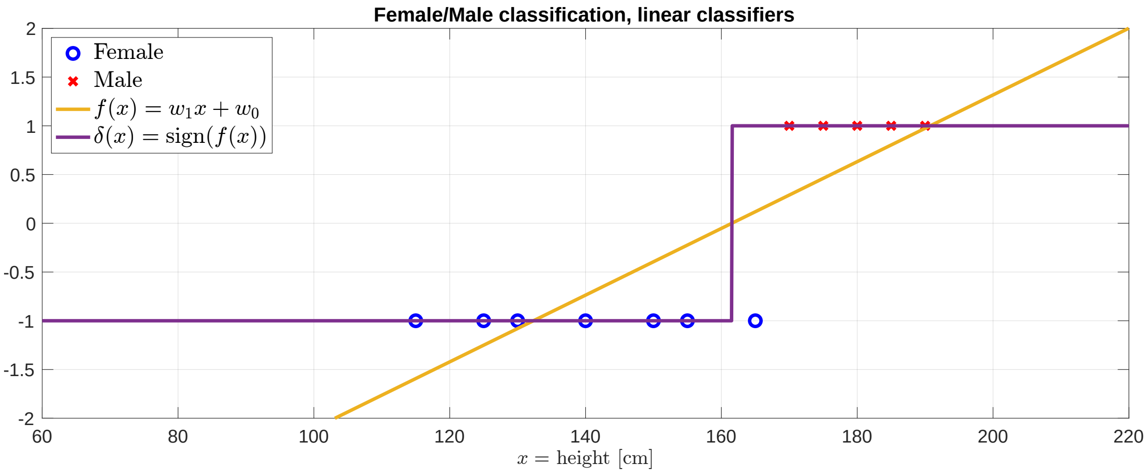
A new point to clasify: $x^Q = 163$

Which class does x^Q belong to? $d^Q = ?$

Linear function LSQ fit



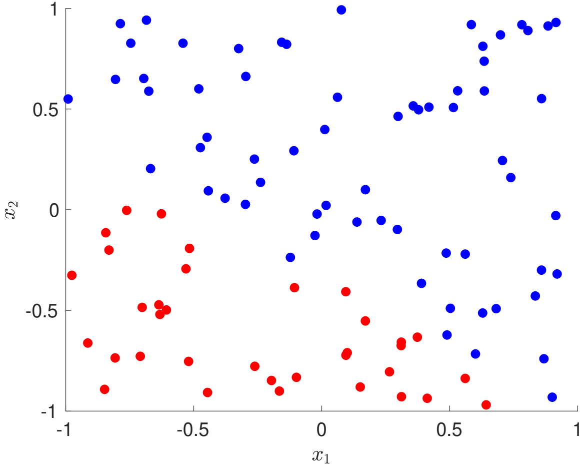
Linear function LSQ fit, discriminant function



Can we do better than fitting a linear function?

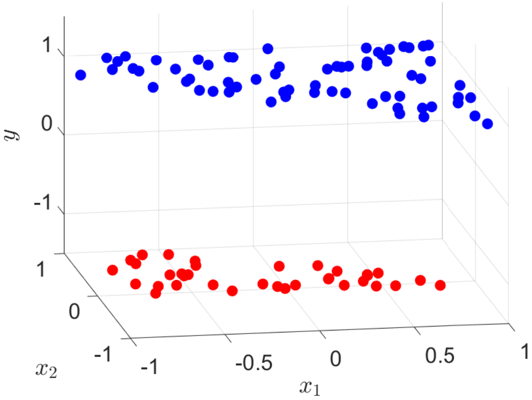
Recap the naive linear approach first.

Learning linear classifier: naive approach, illustration



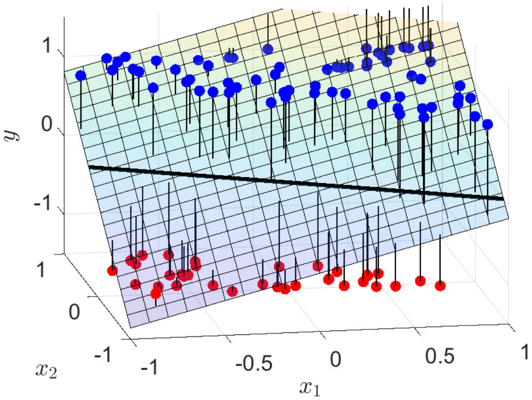
Given a dataset of input vectors $\boldsymbol{x}^{(i)}$ and their classes $y^{(i)}$ ...

Learning linear classifier: naive approach, illustration



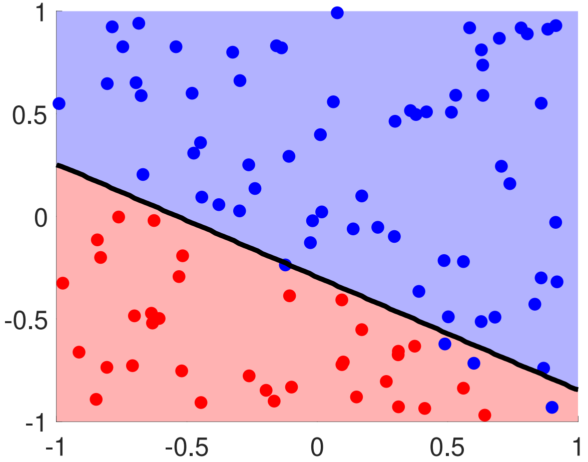
. . . we shall encode the class label as $y = -1$ and $y = 1$ . . .

Learning linear classifier: naive approach, illustration



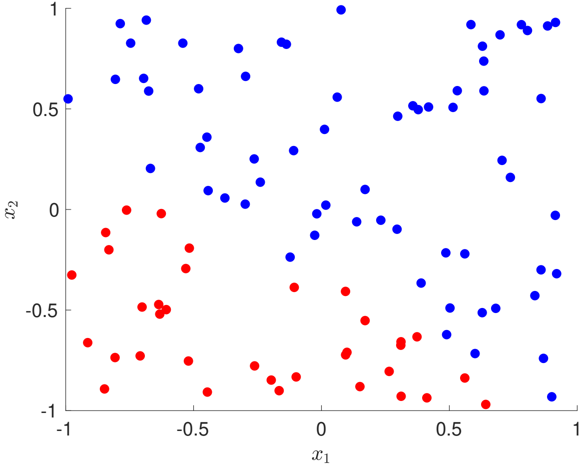
...and fit a linear discriminant function by minimizing MSE as in regression. The contour line $y = 0$ ...

Learning linear classifier: naive approach, illustration



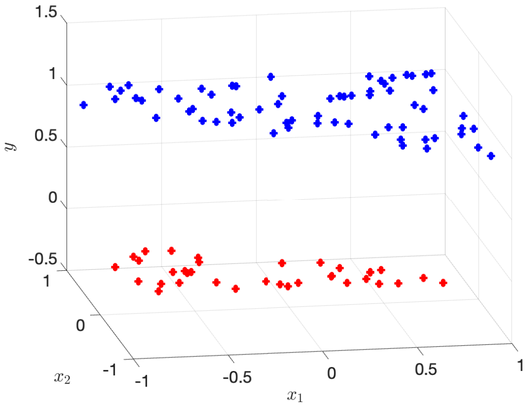
…then forms a linear decision boundary in the original 2D space.
But is such a classifier good in general?

Fitting a better function: Logistic regression



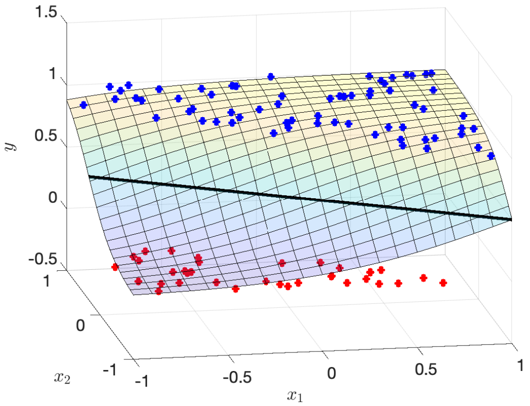
Given a dataset of input vectors $\boldsymbol{x}^{(i)}$ and their classes $y^{(i)}$ ...

Fitting a better function: Logistic regression



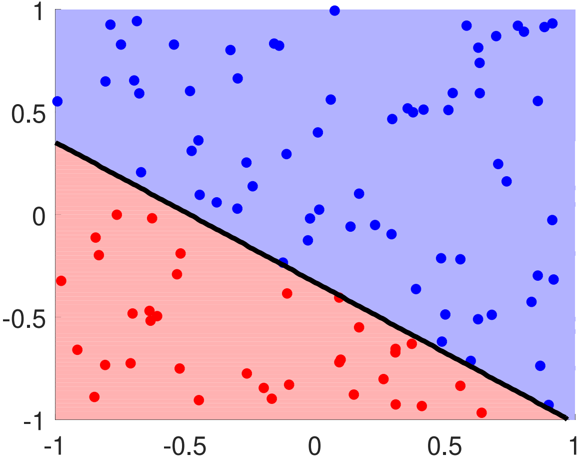
…we shall encode the class label as $y = 0$ and $y = 1$ …

Fitting a better function: Logistic regression



. . . and fit a sigmoidal discriminant function with the threshold 0.5 . . .

Fitting a better function: Logistic regression



. . . which forms a linear decision boundary in the original 2D space.

Logistic regression model

Logistic regression uses a discriminant function which is a nonlinear transformation of the values of a linear function

$$f_{\boldsymbol{w}}(\boldsymbol{x}) = g(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}}},$$

where $g(z) = \dfrac{1}{1 + e^{-z}}$ is the sigmoid function (a.k.a logistic function).

Interpretation of the model:

- $f_{\boldsymbol{w}}(\boldsymbol{x})$ is interpretted as an estimate of the probability that $\boldsymbol{x}$ belongs to class 1.
- The decision boundary is defined using a different level-set: $\{\boldsymbol{x} : f_{\boldsymbol{w}}(\boldsymbol{x}) = 0.5\}$.
- Logistic *regression* is a *classification model!*
- The discriminant function $f_{\boldsymbol{w}}(\boldsymbol{x})$ itself is not linear anymore; but the *decision boundary is still linear!*
- Thanks to the sigmoidal transformation, logistic regression is much less influenced by examples far from the decision boundary!

Logistic regression model

Logistic regression uses a discriminant function which is a nonlinear transformation of the values of a linear function

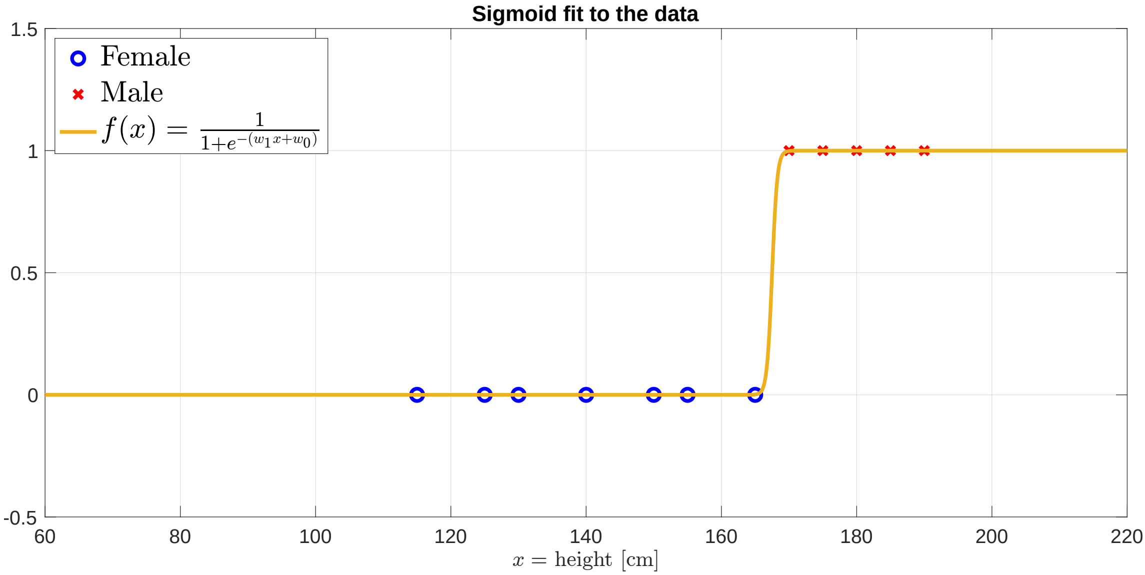
$$f_{\boldsymbol{w}}(\boldsymbol{x}) = g(\boldsymbol{w}^\top \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}}},$$

where $g(z) = \dfrac{1}{1 + e^{-z}}$ is the sigmoid function (a.k.a logistic function).

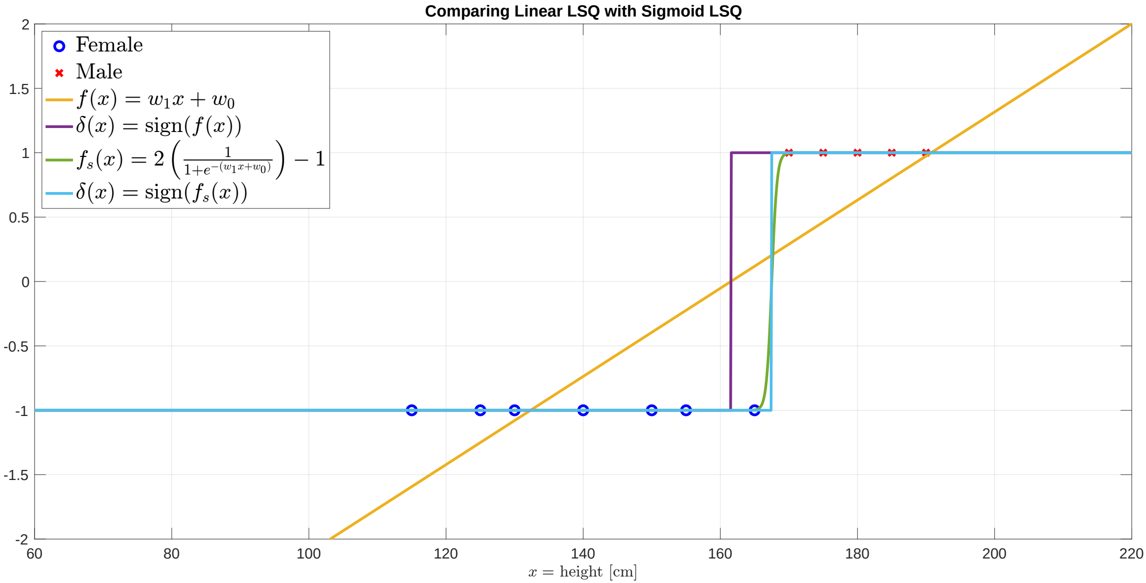
Interpretation of the model:

- $f_{\boldsymbol{w}}(\boldsymbol{x})$ is interpretted as an estimate of the probability that $\boldsymbol{x}$ belongs to class 1.
- The decision boundary is defined using a different level-set: $\{\boldsymbol{x} : f_{\boldsymbol{w}}(\boldsymbol{x}) = 0.5\}$.
- Logistic *regression* is a *classification model!*
- The discriminant function $f_{\boldsymbol{w}}(\boldsymbol{x})$ itself is not linear anymore; but the *decision boundary is still linear!*
- Thanks to the sigmoidal transformation, logistic regression is much less influenced by examples far from the decision boundary!

Sigmoid LSQ fit



Comparing Linear and Sigmoid LSQ fit



What is the proper loss function ℓ?

To train the logistic regression model, one can minimize the J_{MSE} criterion:

- results in a non-convex, multimodal landscape which is hard to optimize.

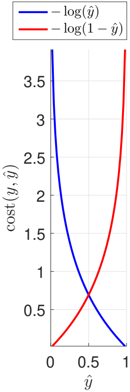
Log. reg. uses a loss function called "cross-entropy":

$$J(\boldsymbol{w}, T) = \frac{1}{N}\sum_{i=1}^{N} \ell(y^{(i)}, f_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})), \text{ where}$$

$$\ell(y, \widehat{y}) = \begin{cases} -\log(\widehat{y}) & \text{if } y = 1 \\ -\log(1-\widehat{y}) & \text{if } y = 0 \end{cases},$$

which can be rewritten in a single expression as

$$\ell(y, \widehat{y}) = -y \cdot \log(\widehat{y}) - (1-y) \cdot \log(1-\widehat{y}).$$



- simpler to optimize for numerical solvers.

What is the proper loss function ℓ?

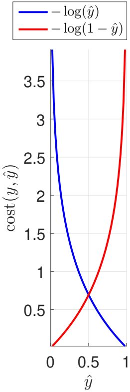
To train the logistic regression model, one can minimize the J_{MSE} criterion:

- results in a non-convex, multimodal landscape which is hard to optimize.

Log. reg. uses a loss function called cross-entropy :

$$J(\boldsymbol{w}, \mathcal{T}) = \frac{1}{N} \sum_{i=1}^{N} \ell(y^{(i)}, f_{\boldsymbol{w}}(\boldsymbol{x}^{(i)})), \text{ where}$$

$$\ell(y, \widehat{y}) = \begin{cases} -\log(\widehat{y}) & \text{if } y = 1 \\ -\log(1 - \widehat{y}) & \text{if } y = 0 \end{cases},$$

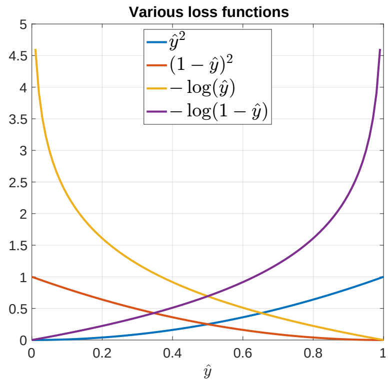


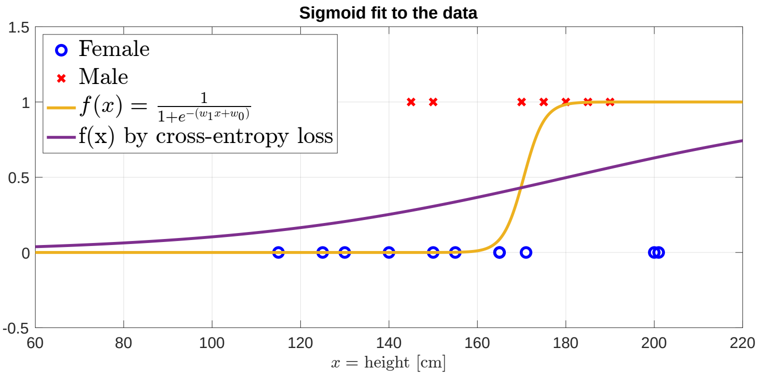
which can be rewritten in a single expression as

$$\ell(y, \widehat{y}) = -y \cdot \log(\widehat{y}) - (1 - y) \cdot \log(1 - \widehat{y}).$$

- simpler to optimize for numerical solvers.

MSE vs cross entropy loss



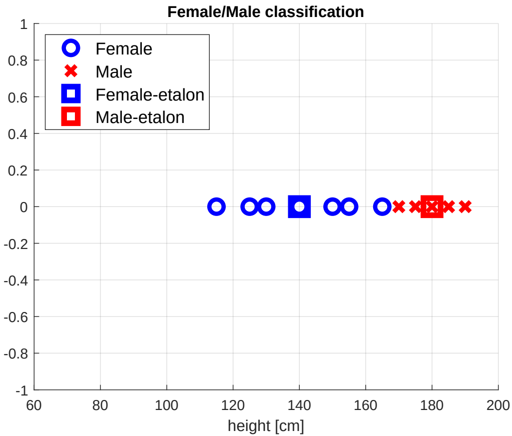


Sigmoidal $f(x)$ can be also interpreted as $p(s = \text{Male} \mid x)$ – Learning Dicriminative model directly.

Cross-entropy loss strongly penalizes hard errors, complete mismatches.

Alternative idea: F/M classification – Etalons

Represent each class by a single example called *etalon*! (Or by a very small number of etalons.)



$$e_F = \mathrm{ave}(\{x^{(i)} : s^{(i)} = F\}) = 140$$
$$e_M = \mathrm{ave}(\{x^{(i)} : s^{(i)} = M\}) = 180$$

Based on etalons: $d_Q = ?$

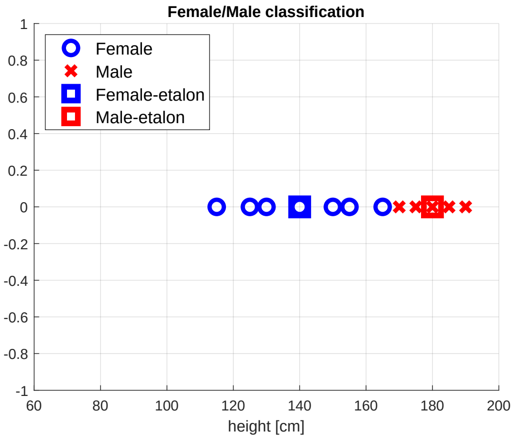
- A $d^Q = F$
- B $d_Q = M$
- C Both classes equally likely
- D Cannot provide any decision

Classify as $d^Q = \mathrm{argmin}_{s \in S} \mathrm{dist}(x^Q, e_s)$

What type of function is $\mathrm{dist}(x^Q, e_s)$?

Alternative idea: F/M classification – Etalons

Represent each class by a single example called *etalon*! (Or by a very small number of etalons.)



$$e_F = \text{ave}(\{x^{(i)} : s^{(i)} = F\}) = 140$$
$$e_M = \text{ave}(\{x^{(i)} : s^{(i)} = M\}) = 180$$

Based on etalons: $d_Q = ?$

A $d^Q = F$

B $d_Q = M$

C Both classes equally likely

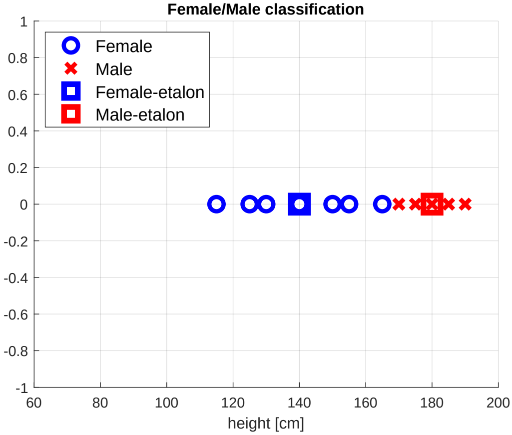
D Cannot provide any decision

Classify as $d^Q = \text{argmin}_{s \in S} \text{dist}(x^Q, e_s)$

What type of function is $\text{dist}(x^Q, e_s)$?

Alternative idea: F/M classification – Etalons

Represent each class by a single example called *etalon*! (Or by a very small number of etalons.)



$$e_F = \text{ave}(\{x^{(i)} : s^{(i)} = F\}) = 140$$
$$e_M = \text{ave}(\{x^{(i)} : s^{(i)} = M\}) = 180$$

Based on etalons: $d_Q = ?$

A $d^Q = F$

B $d_Q = M$

C Both classes equally likely

D Cannot provide any decision

Classify as $d^Q = \text{argmin}_{s \in \mathcal{S}} \text{dist}(x^Q, e_s)$

What type of function is $\text{dist}(x^Q, e_s)$?

Etalon classifier is a Linear classifier

Assuming $\text{dist}(x, e) = (x - e)^2$, then

$$\operatorname*{argmin}_{s \in S} \text{dist}(x, e_s) = \operatorname*{argmin}_{s \in S} (x - e_s)^2 = \operatorname*{argmin}_{s \in S} (\underbrace{x^2}_{\text{const.}} - 2e_s x + e_s^2) =$$

$$= \operatorname*{argmin}_{s \in S} (-2e_s x + e_s^2) = \operatorname*{argmax}_{s \in S} (\underbrace{e_s x - \frac{1}{2} e_s^2}_{\text{linear function of } x})$$

Multiclass classification: each class s has a linear discriminant function $f_s(x) = a_s x + b_s$ and

$$\delta(x) = \operatorname*{argmax}_{s \in S} f_s(x)$$

Binary classification: a single linear discriminant function $g(x)$ is sufficient and

$$\delta(x) = \begin{cases} s_1 & \text{if } g(x) \geq 0 \\ s_2 & \text{if } g(x) < 0 \end{cases}$$

Etalon classifier is a Linear classifier

Assuming $\mathrm{dist}(x, e) = (x - e)^2$, then

$$\operatorname*{argmin}_{s \in S} \mathrm{dist}(x, e_s) = \operatorname*{argmin}_{s \in S} (x - e_s)^2 = \operatorname*{argmin}_{s \in S} (\underbrace{x^2}_{\text{const.}} - 2e_s x + e_s^2) =$$

$$= \operatorname*{argmin}_{s \in S} (-2e_s x + e_s^2) = \operatorname*{argmax}_{s \in S} (\underbrace{e_s x - \frac{1}{2} e_s^2}_{\text{linear function of } x})$$

Multiclass classification: each class s has a linear discriminant function $f_s(x) = a_s x + b_s$ and

$$\delta(x) = \operatorname*{argmax}_{s \in S} f_s(x)$$

Binary classification: a single linear discriminant function $g(x)$ is sufficient and

$$\delta(x) = \begin{cases} s_1 & \text{if } g(x) \geq 0 \\ s_2 & \text{if } g(x) < 0 \end{cases}$$

Etalon classifier is a Linear classifier

Assuming $\text{dist}(x, e) = (x - e)^2$, then

$$\underset{s\in S}{\text{argmin}}\,\text{dist}(x, e_s) = \underset{s\in S}{\text{argmin}}(x - e_s)^2 = \underset{s\in S}{\text{argmin}}(\underbrace{x^2}_{\text{const.}} - 2e_s x + e_s^2) =$$

$$= \underset{s\in S}{\text{argmin}}(-2e_s x + e_s^2) = \underset{s\in S}{\text{argmax}}(\underbrace{e_s x - \frac{1}{2}e_s^2}_{\text{linear function of } x})$$

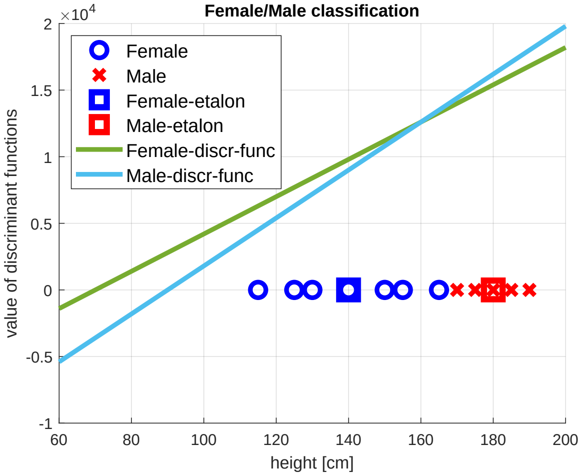
Multiclass classification: each class s has a linear discriminant function $f_s(x) = a_s x + b_s$ and

$$\delta(x) = \underset{s\in S}{\text{argmax}}\, f_s(x)$$

Binary classification: a single linear discriminant function $g(x)$ is sufficient and

$$\delta(x) = \begin{cases} s_1 & \text{if } g(x) \geq 0 \\ s_2 & \text{if } g(x) < 0 \end{cases}$$

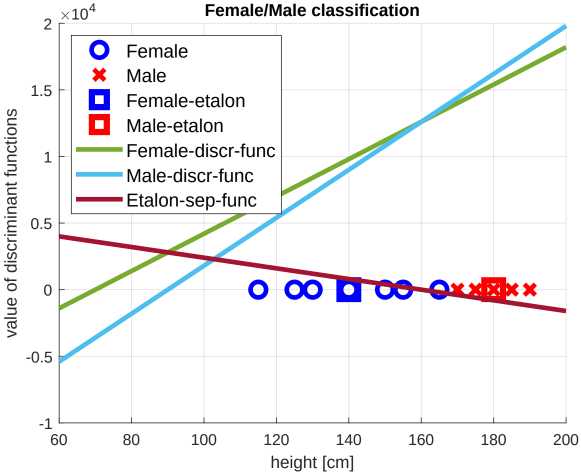
Example: F/M – Linear discriminant functions based on etalons



Discriminant functions for 2 classes:

$$f_F(x) = a_F x + b_F =$$
$$= e_F x - \frac{1}{2} e_F^2 = 140x - 9800$$
$$f_M(x) = a_M x + b_M =$$
$$= e_M x - \frac{1}{2} e_M^2 = 180x - 16200$$

Example: F/M – Linear discriminant functions based on etalons



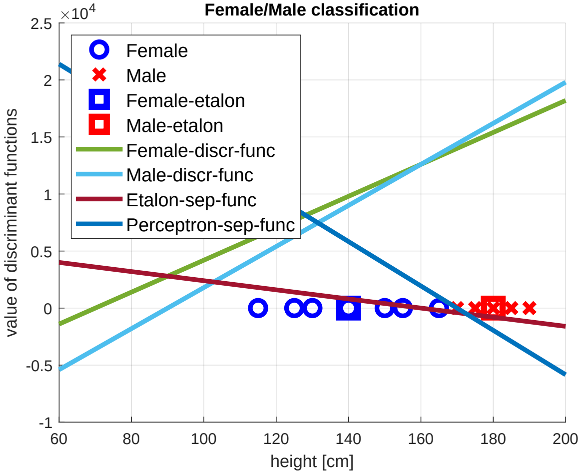
Discriminant functions for 2 classes:

$$f_F(x) = a_F x + b_F = $$
$$= e_F x - \frac{1}{2} e_F^2 = 140x - 9800$$
$$f_M(x) = a_M x + b_M = $$
$$= e_M x - \frac{1}{2} e_M^2 = 180x - 16200$$

A single discriminant function separating 2 classes:

$$g(x) = f_F(x) - f_M(x) = $$
$$= -40x + 6400$$

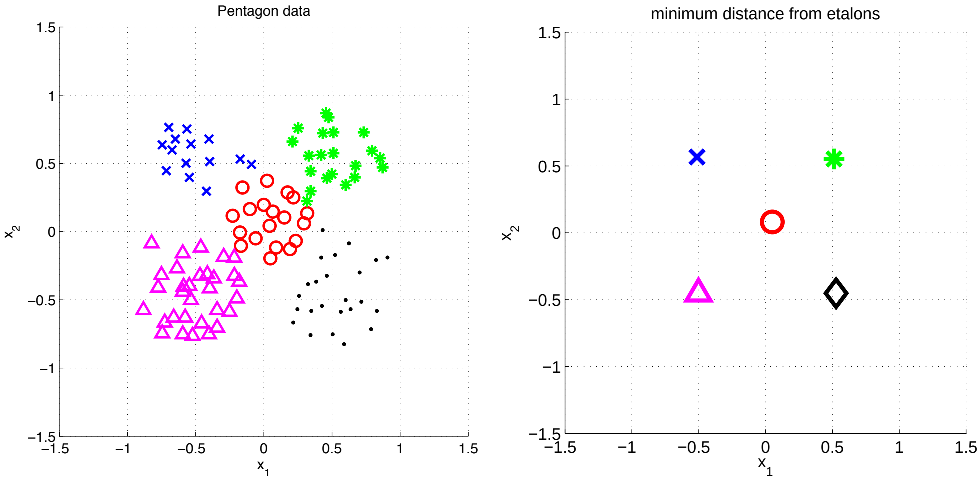
Example: F/M – Can we do better etalons?



Etalon-based linear classifier makes some errors.

A perceptron algorithm may be used to find a zero-error classifier (if one exists).

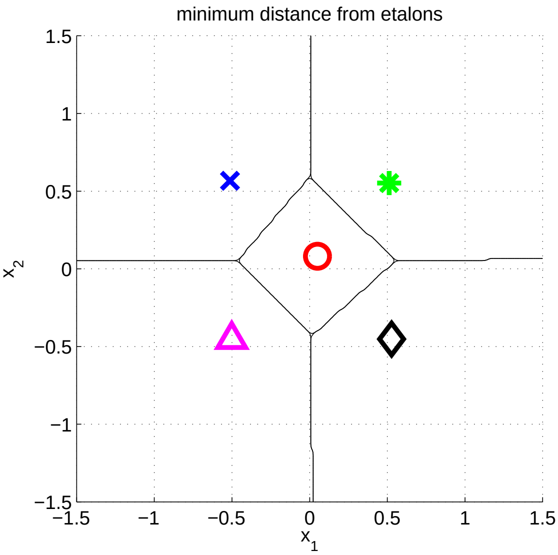
Etalon based classification



Represent $\vec{x}$ by etalon , $\vec{e}_s$ per each class $s \in S$.

Separate etalons

$$s^* = \arg\min_{s \in S} \|\vec{x} - \vec{e}_s\|^2$$

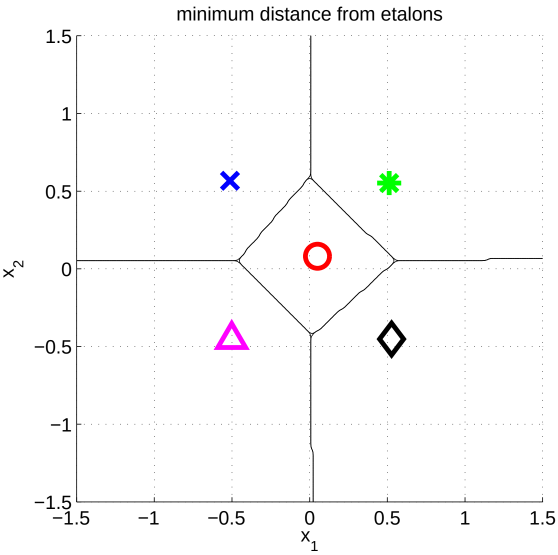


What etalons?

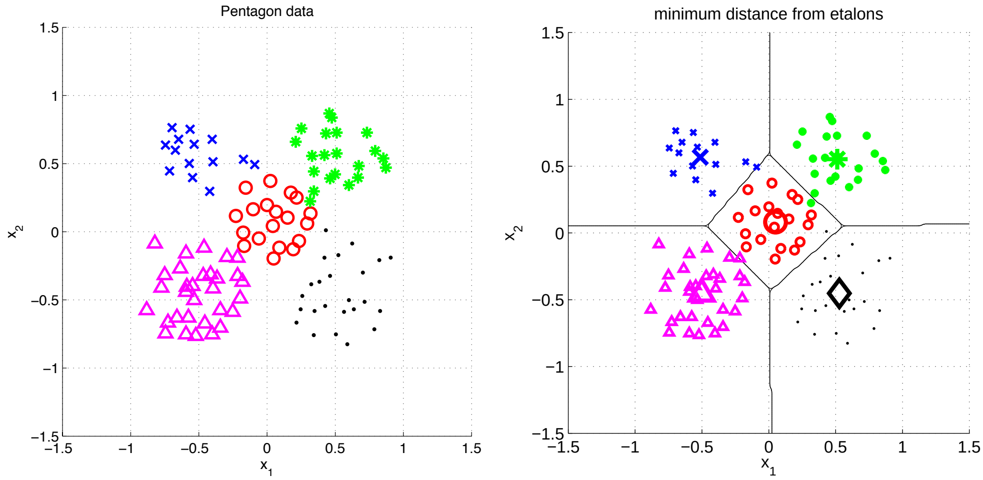
If $\mathcal{N}(\vec{x}|\vec{\mu},\Sigma)$; all classes same covariance matrices, then

$$\vec{e}_s \stackrel{\text{def}}{=} \vec{\mu}_s = \frac{1}{|\mathcal{X}^s|}\sum_{i\in\mathcal{X}^s} \vec{x}_i^s$$

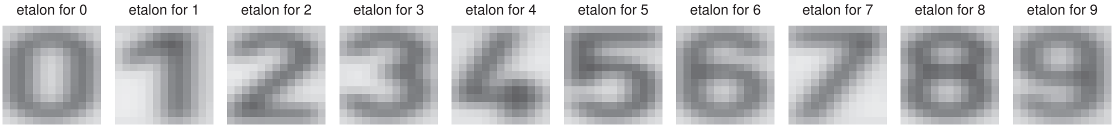
and separating hyperplanes halve distances between pairs.



Etalon based classification, $\vec{e}_s = \vec{\mu}_s$



Digit recognition - etalons $\vec{e}_s = \vec{\mu}_s$



Figures from [7].

Bayesian Discriminant functions $f(\vec{x}, s)$, $g_s(\vec{x})$

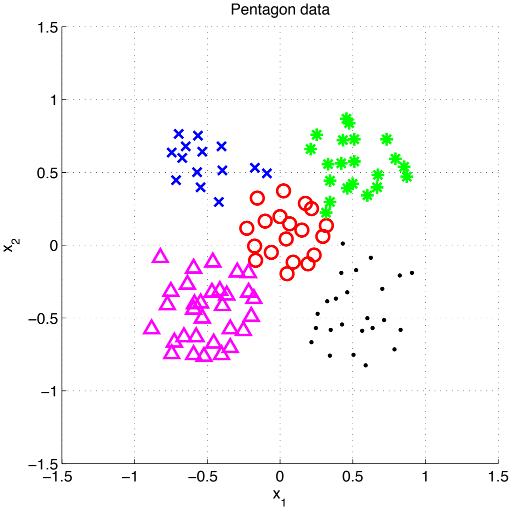
$$s^* = \operatorname*{argmax}_{s \in \mathcal{S}} f(\vec{x}, s)$$

Bayes:

$$s^* = \operatorname*{argmax}_{s \in \mathcal{S}} P(s|\vec{x}) = \frac{P(\vec{x} \mid s)P(s)}{P(\vec{x})}$$

Discriminant function:

$$f(\vec{x}, s) = g_s(\vec{x}) = P(\vec{x} \mid s)P(s)$$



Etalon classifier – Linear classifier, generalization to higher dimensions

$$
\begin{aligned}
s^* &= \arg\min_{s \in S} \|\vec{x} - \vec{e}_s\|^2 = \arg\min_{s \in S} (\vec{x}^\top \vec{x} - 2\,\vec{e}_s^\top \vec{x} + \vec{e}_s^\top \vec{e}_s) = \\
&= \arg\min_{s \in S} \left(\vec{x}^\top \vec{x} - 2\left(\vec{e}_s^\top \vec{x} - \frac{1}{2}(\vec{e}_s^\top \vec{e}_s)\right)\right) = \\
&= \arg\min_{s \in S} \left(\vec{x}^\top \vec{x} - 2\left(\vec{e}_s^\top \vec{x} + b_s\right)\right) = \\
&= \boxed{\arg\max_{s \in S} (\vec{e}_s^\top \vec{x} + b_s)} = \arg\max_{s \in S} g_s(\vec{x}). \qquad b_s = -\frac{1}{2}\vec{e}_s^\top \vec{e}_s
\end{aligned}
$$

Linear function (plus offset)

$$g_s(\mathbf{x}) = \mathbf{w}_s^\top \mathbf{x} + w_{s0}$$

Learning and decision

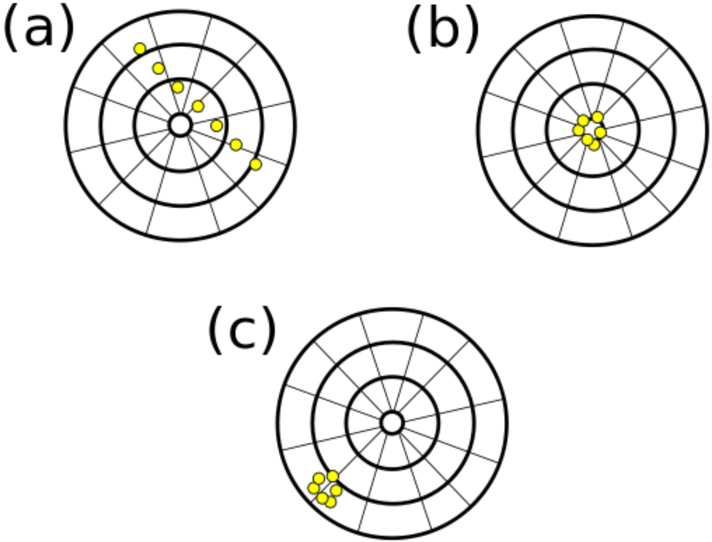
Learning stage - learning models/function/parameters from data.

Decision stage - decide about a query $\vec{x}$.

What to learn?

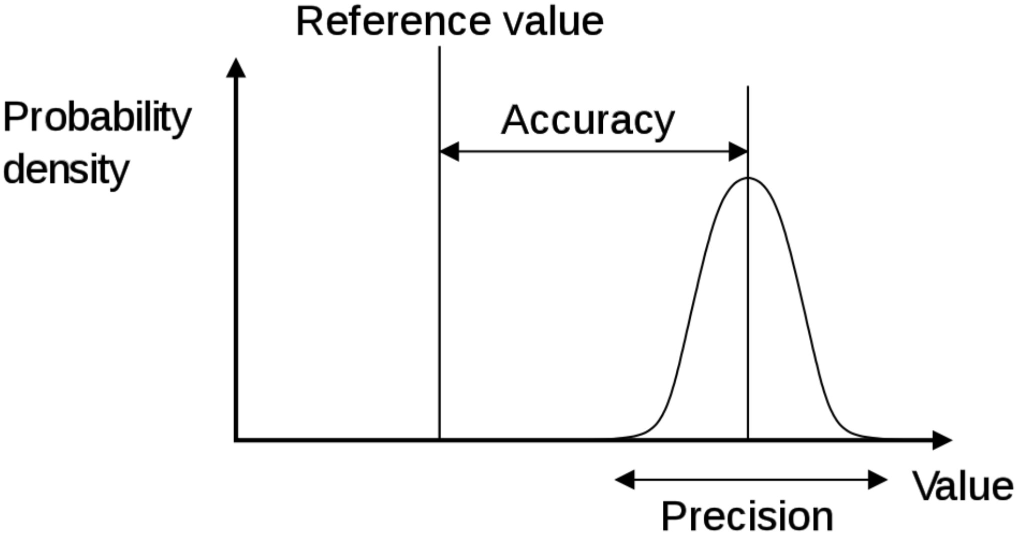
- Generative model : Learn $P(\vec{x}, s)$. Decide by computing $P(s|\vec{x})$.
- Discriminative model : Learn $P(s|\vec{x})$.
- Discriminant function : Learn $g(\vec{x})$ which maps $\vec{x}$ directly into class labels.

Accuracy vs precision



https://commons.wikimedia.org/wiki/File:Precision_versus_accuracy.svg

Accuracy vs precision



https://en.wikipedia.org/wiki/Accuracy_and_precision

References I

Further reading: Chapter 18 of [6], or chapter 4 of [1], or chapter 5 of [2]. Many figures created with the help of [3]. You may also play with demo functions from [7].
Human deciding and predicting under noise, [4] (in Czech [5])

[1] Christopher M. Bishop.
Pattern Recognition and Machine Learning.
Springer Science+Bussiness Media, New York, NY, 2006.
https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork.
Pattern Classification.
John Wiley & Sons, 2nd edition, 2001.

[3] Vojtěch Franc and Václav Hlaváč.
Statistical pattern recognition toolbox.
http://cmp.felk.cvut.cz/cmp/software/stprtool/index.html.

References II

[4] D. Kahneman, O. Sibony, and C.R. Sunstein.
Noise: A Flaw in Human Judgment.
Little Brown Spark, 2021.

[5] D. Kahneman, O. Sibony, and C.R. Sunstein.
Šum, O chybách v lidském úsudku.
Jan Melvil Publishing, 2021.

[6] Stuart Russell and Peter Norvig.
Artificial Intelligence: A Modern Approach.
Prentice Hall, 3rd edition, 2010.
http://aima.cs.berkeley.edu/.

[7] Tomáš Svoboda, Jan Kybic, and Hlaváč Václav.
Image Processing, Analysis and Machine Vision — A MATLAB Companion.
Thomson, Toronto, Canada, 1st edition, September 2007.
http://visionbook.felk.cvut.cz/.