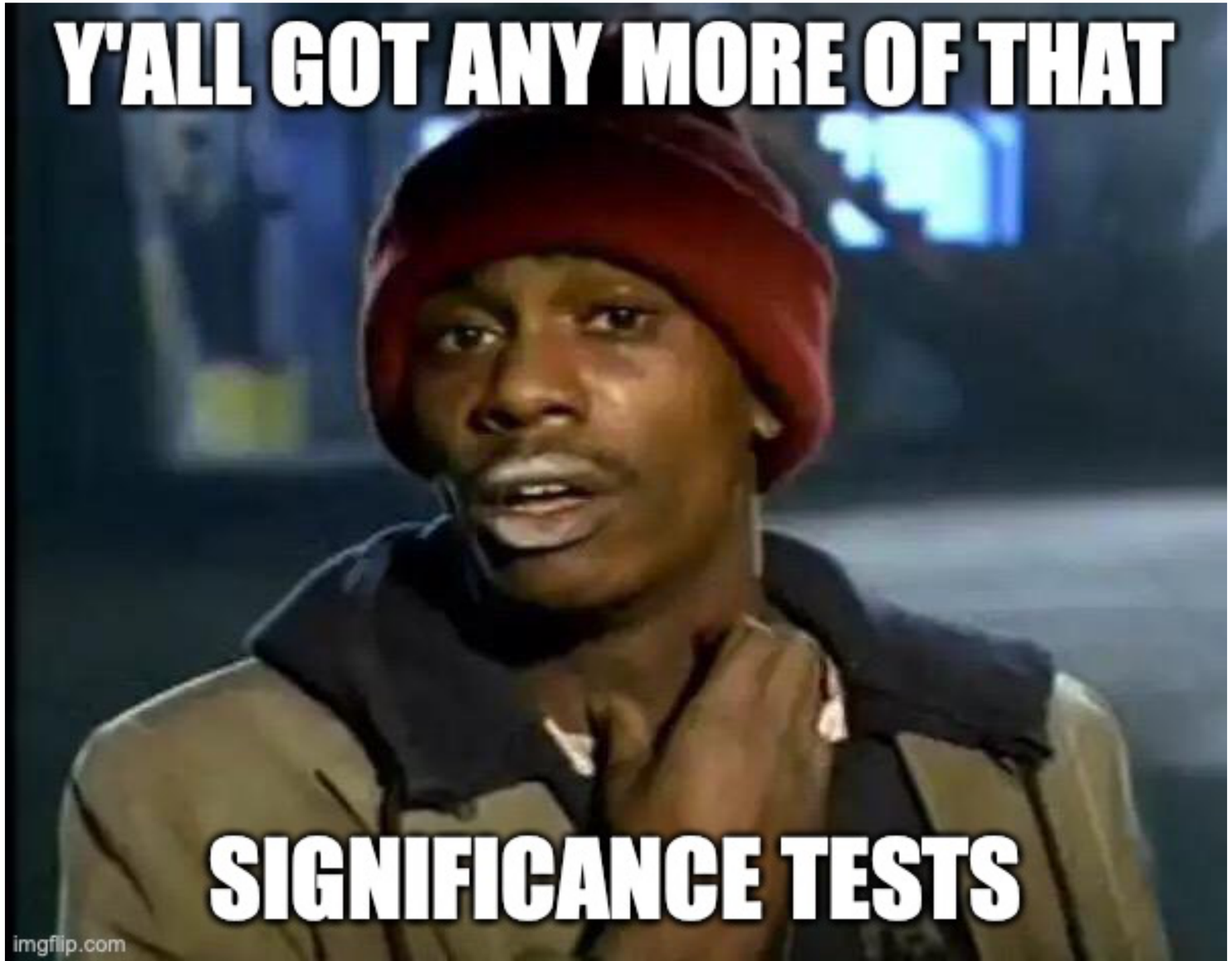
Statistical Techniques for Data Science & Robotics

Week 4

Objectives (for today)

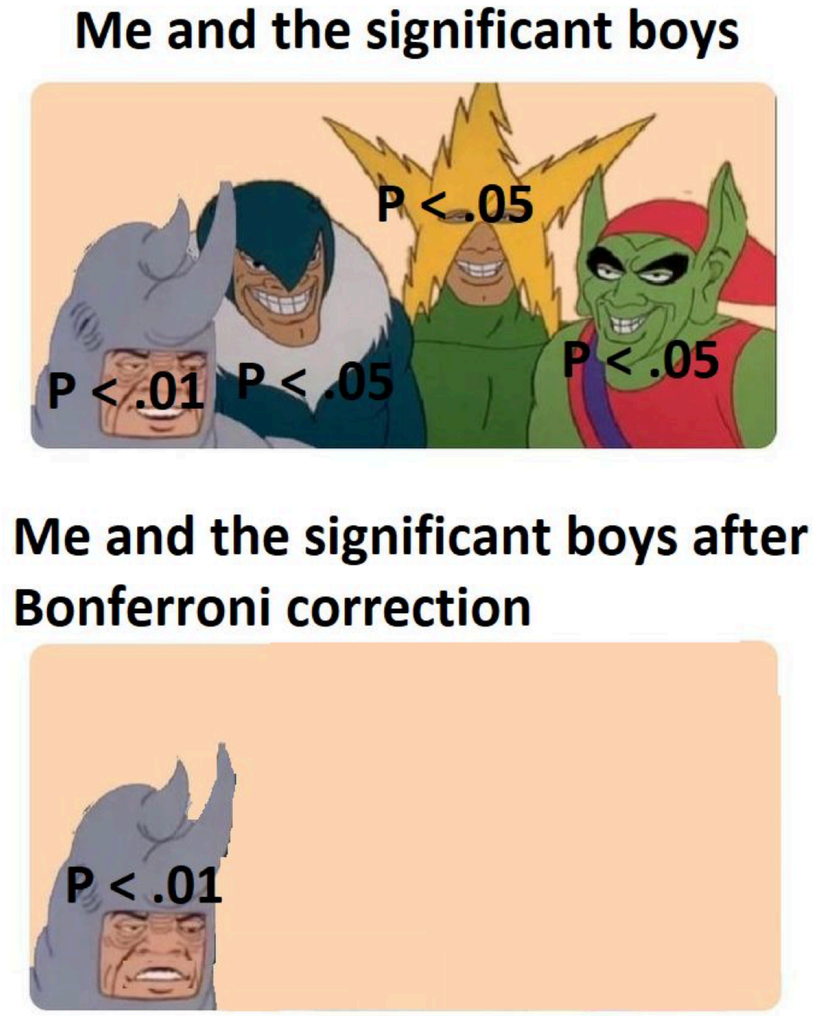
- Recap of the last week
- Decision Theory / Classification
- Regression Models:
  - Loss function for regression
  - Linear Regression
  - The Bias-Variance Decomposition

Last week debt / recap



Multiple testing and Correction

- Significance tests:
  - Z-test,
  - t-test,
  - Chi-square GoF
  - ANOVA



- Bonferroni Correction

when you do multiple testing, adjust the significance level, α

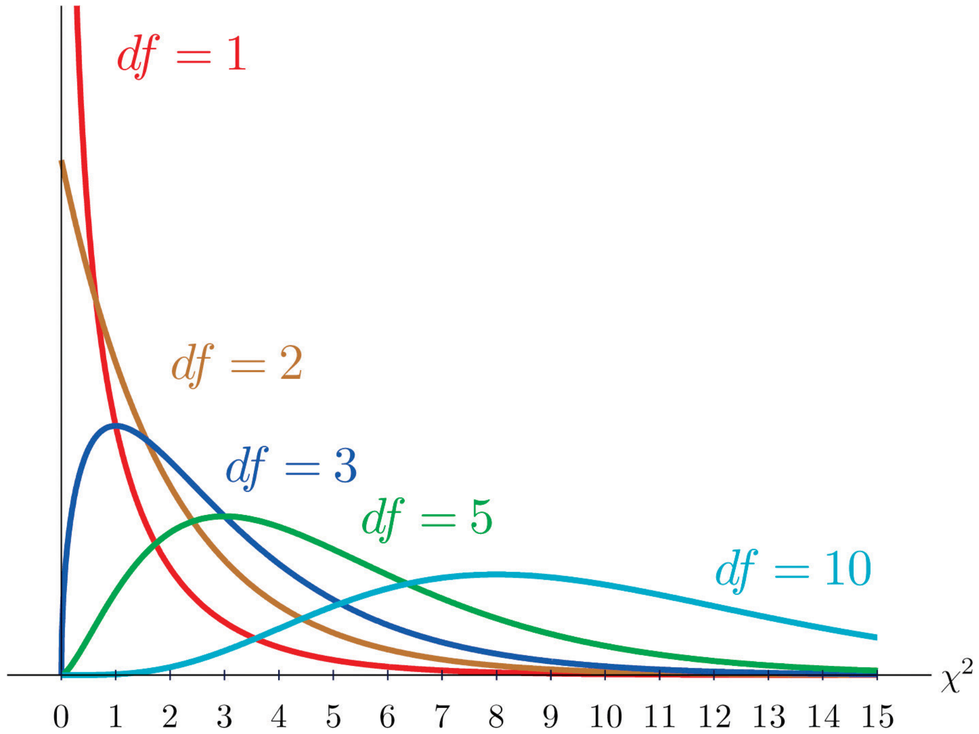
$$\alpha_{corr} = \frac{\alpha}{\#\ \text{of tests}}$$

Chi-square distribution, χ^2

X_i $(i = 1,\ldots,n)$ is an IID sample

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}$$

this random variable has the chi-square distribution with **n degerees of freedom**



Degrees of freedom, χ^2

X_i $(i = 1,\ldots,n)$ is an IID sample

- Question: what is the disrtibution of $\dfrac{(n-1)S^2}{\sigma^2}$?

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2$$
$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})$$
$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2.$$

χ^2_n χ^2_{n-1} χ^2_1

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

Theorem: If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{\sigma^2}$$

has a chi-squared distribution with $v = n-1$ degrees of freedom.

Decision Theory

Introduction to Decision Theory

- Decision Theory
  - helps us making optimal decisions
  - in case of uncertainty (that is modeled with Prob.Theory)

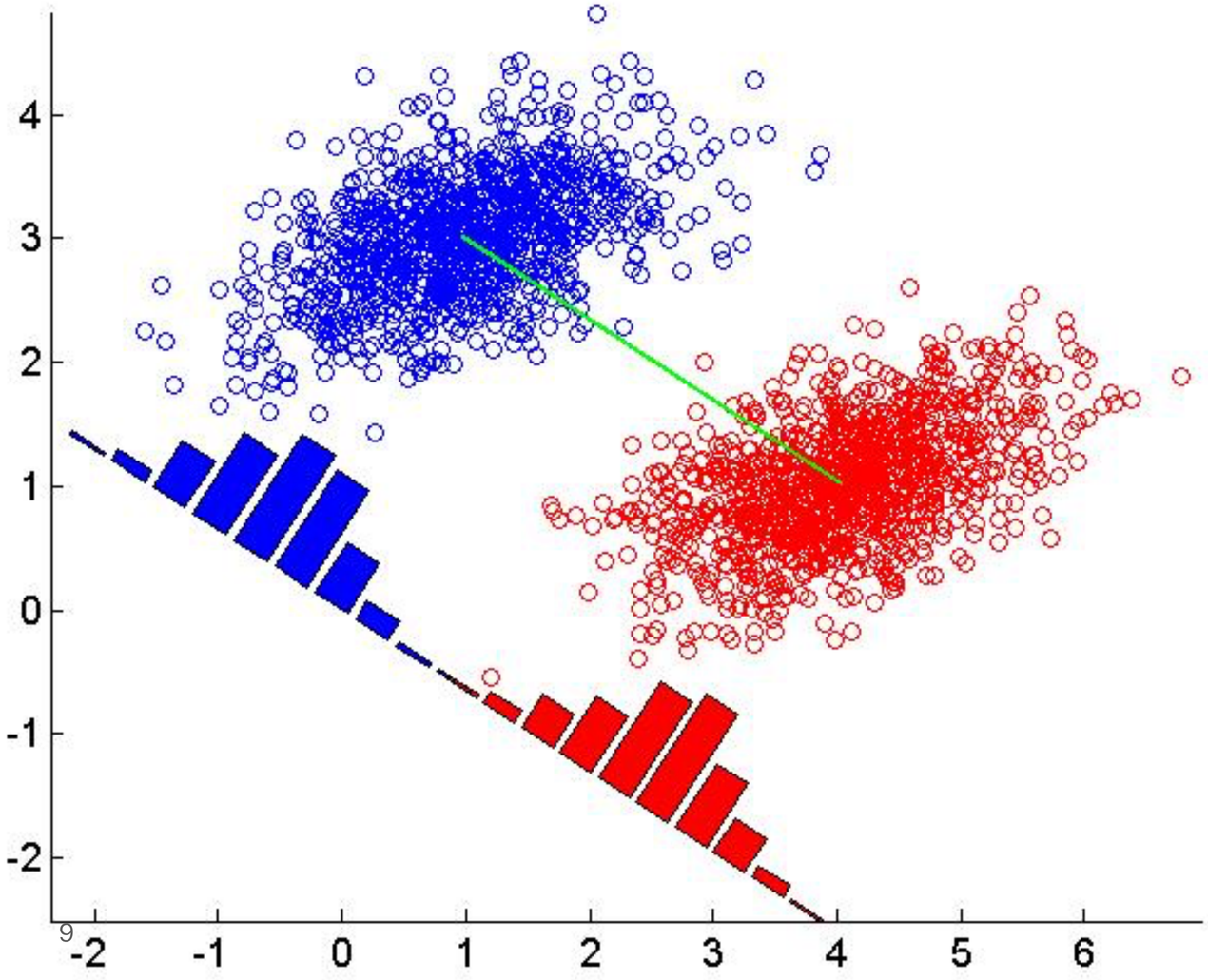
- **First step**: to infer (or learn) the p(x,t) - a joint distribution of inputs and targets
- **Second step**: use probabilities to make optimal decisions

Setup: classification with 2 classes

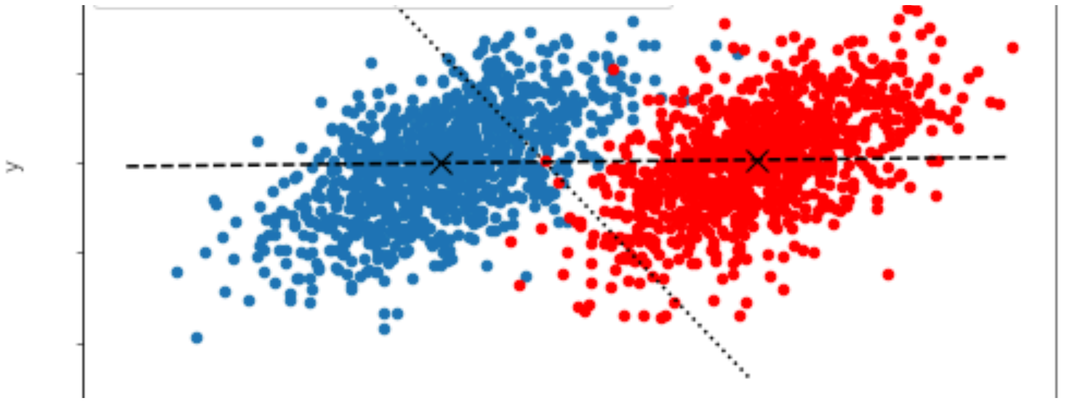
a medical diagnosis problem

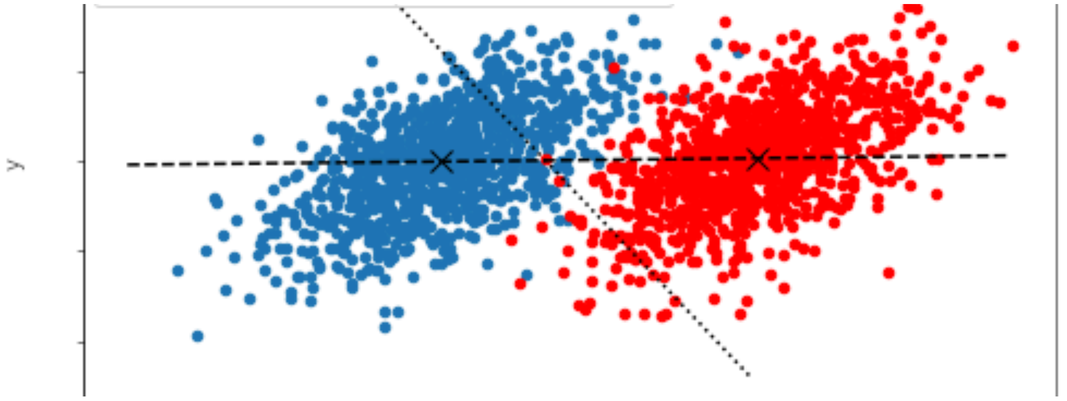
- We need a decision rule:

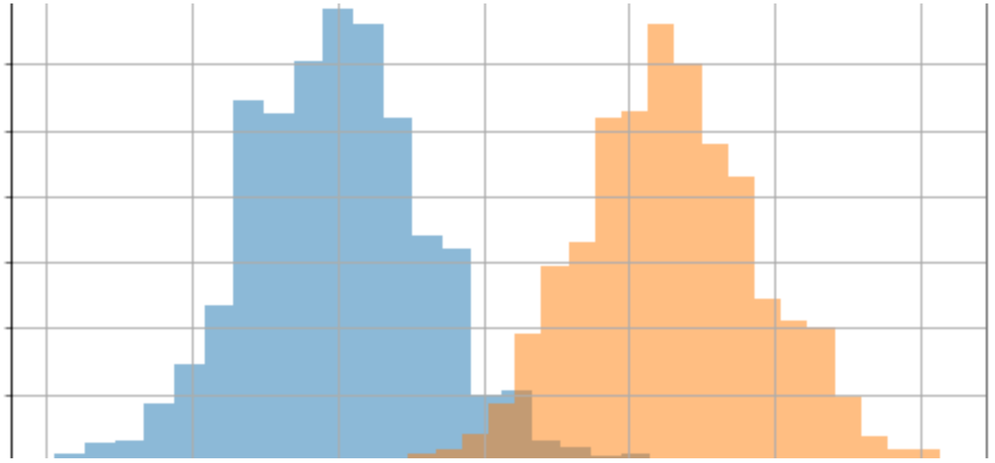
- $\mathbf{x}$ is an input (a D-dim vector);
- $t \in \{0,1\}$;
- $t = 0 \implies C_1$;
- $t = 1 \implies C_2$

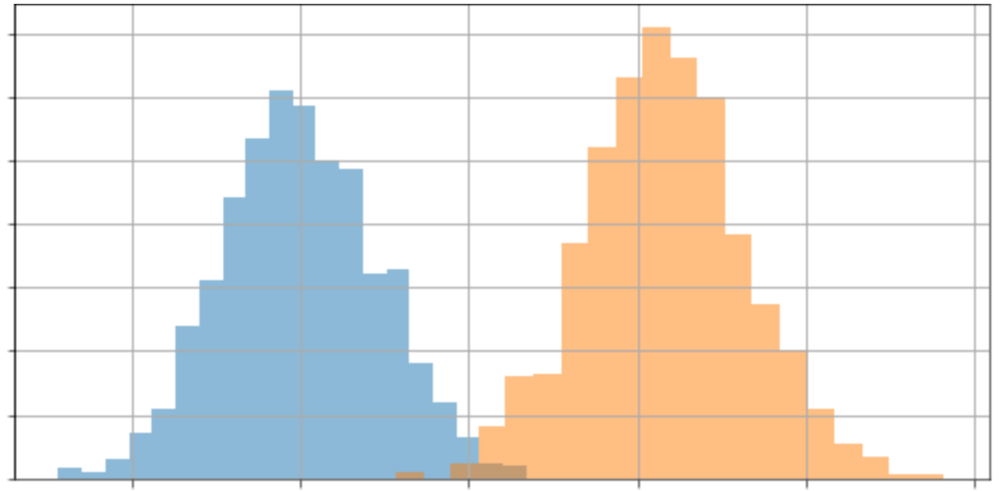


View of data affects the decision









Decision is based on the distribution

a medical diagnosis problem

- => we need a decision rule.
- Setup:
  - $\mathbf{x}$ is an input (a vector);
  - $t \in \{0,1\}$;
  - $t = 0 \implies C_1$;
  - $t = 1 \implies C_2$

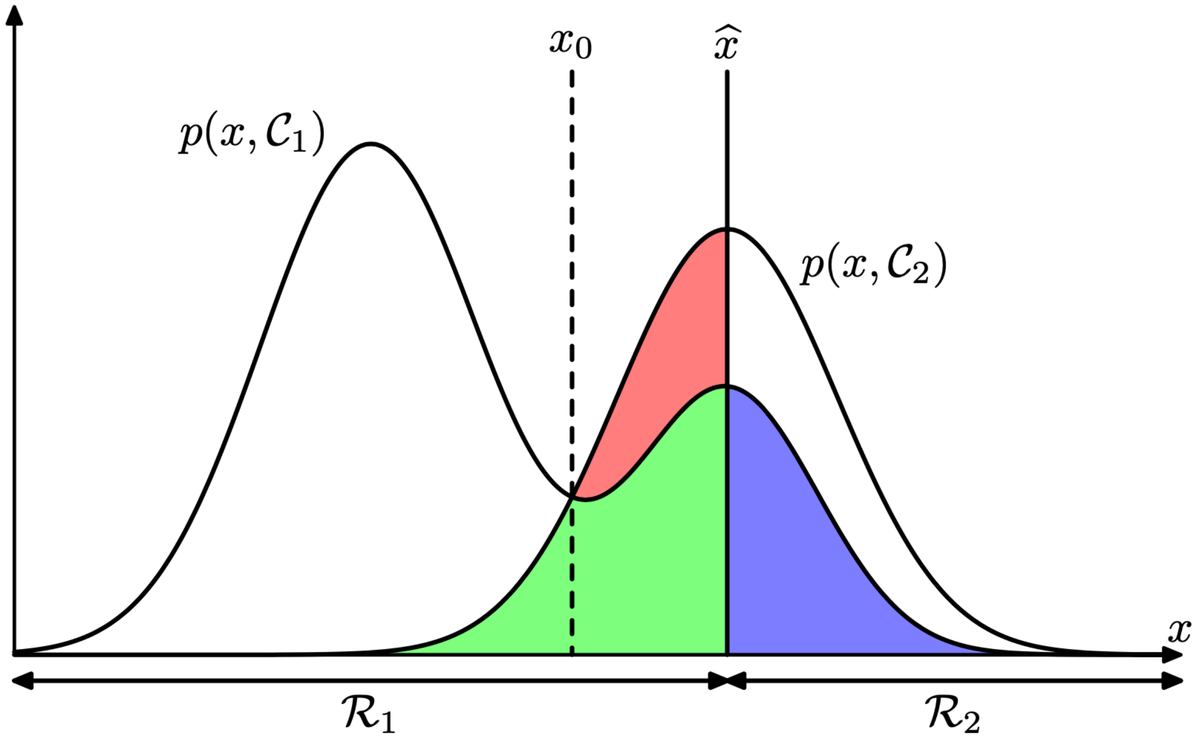
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

k=1,2

Why is the p(x,t) not present here?

Decision regions and boundaries

- R_1, R_2 are decision regions with assignments of **x** to C_1, C_2



- Is $\hat{x}$ the optimal decision boundary?

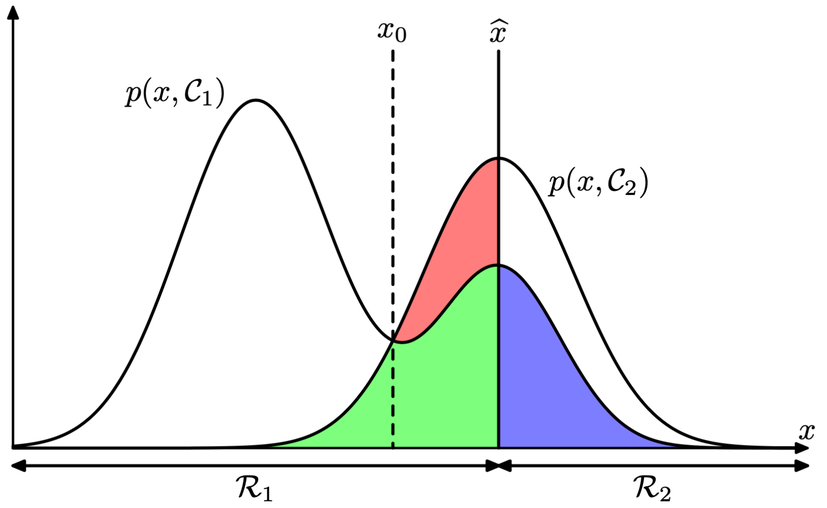
Minimizing errors of classification

- our rule should minimize by picking a decision boundary (x_0):

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}. \end{aligned}$$

- R_1, R_2 are regions with

wrong assignments of **x** to C_2, C_1



Inference and decision

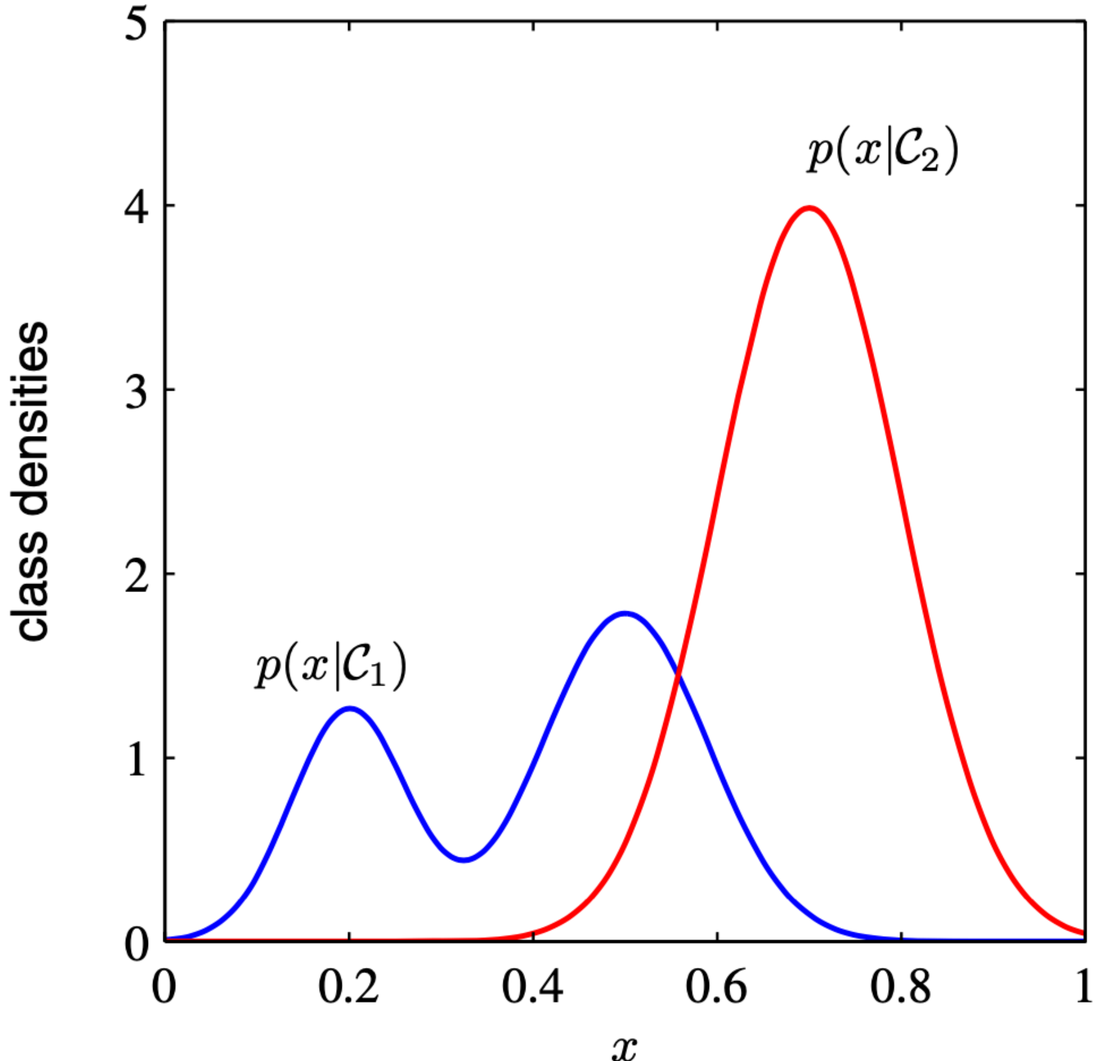
3 **distinct** approaches to solving decision problems

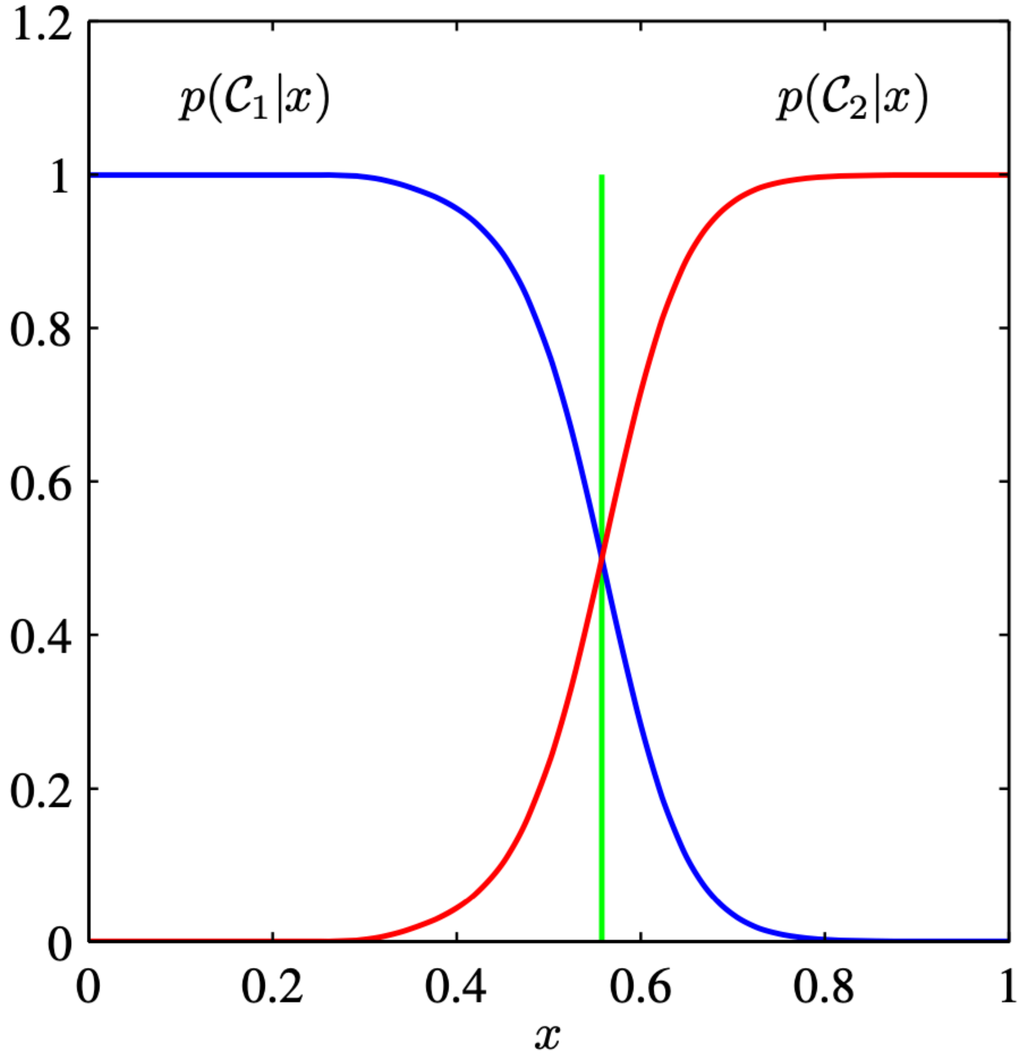
- Determine **class-conditional densities** $p(x \mid C_k)$ for each class C_k, then make a decision for the given x using the formula

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

  - We get a **generative model**
- Determine **posterior class probabilities** $p(C_k \mid x)$ and predict class for an x
  - This is a **discriminative model**
- Find a **discriminant** function, $f(x)$, that maps x to class labels (e.g. a linear clasifier)

Example


class densities
$p(x|\mathcal{C}_2)$
$p(x|\mathcal{C}_1)$
x


$p(\mathcal{C}_1|x)$
$p(\mathcal{C}_2|x)$
x

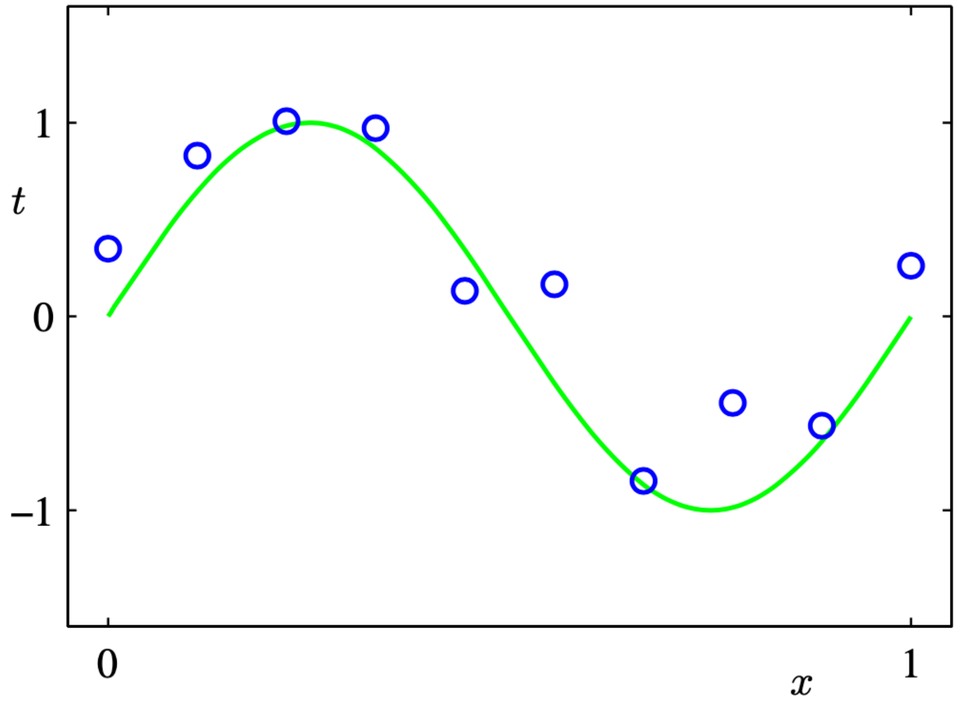
Break, 5 min.

Regression models

Loss functions for regression

Setup

- t is target
- **x** is input (D-dimensional vector)
- y(x) estimate of t for each x
- L(t, y(x)) is a Loss of the estimate
  - common loss: L(t, y(x)) = $(y(x) - t)^2$
- Goal is to Minimize the expected loss



Minimizing Expected Loss

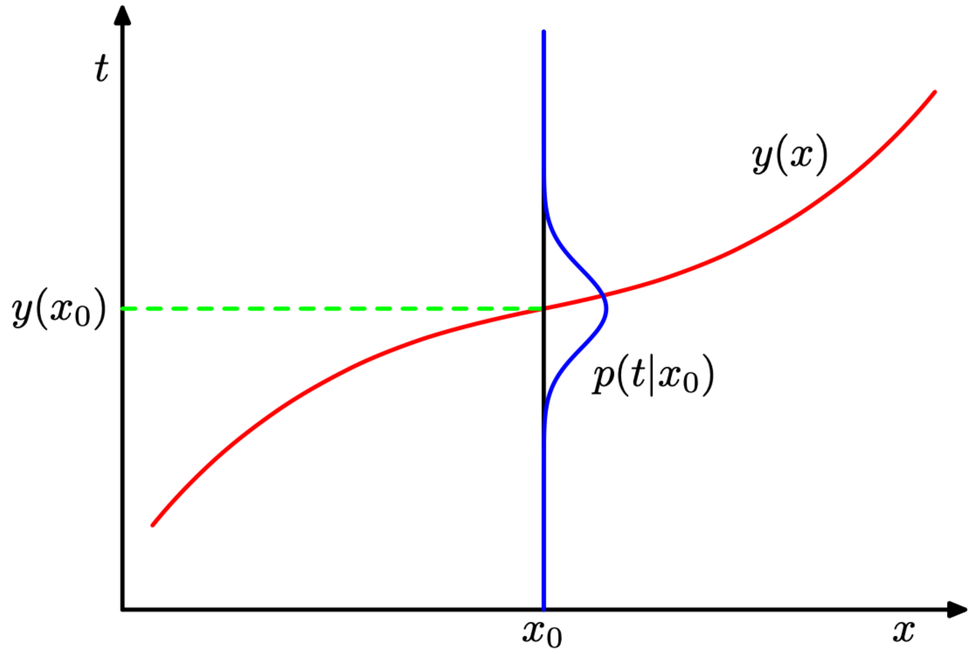
- $\mathbb{E}[L] = \int\int (y(x) - t)^2 p(x,t) \mathrm{dx}\ \mathrm{dt}$
- this is a funcitional of the loss function L: mean squared error for all possible x, t

- To Solve $\min_{y(x)} \mathbb{E}[L]$ $\quad =>\quad$ $\dfrac{\delta \mathbb{E}[L]}{\delta y(x)} = 2\int (y(x) - t) p(x,t)\ \mathrm{dt} = 0$ (from variational calculus)

- The Solution is: $y(x) = \dfrac{\int t p(x,t)\ \mathrm{dt}}{p(x)} = \int t p(t \mid x)\ \mathrm{dt} = \mathbb{E}_t[t \mid x]$
- So, the optimal regression function is the **conditional expectation** $\mathbb{E}_t[t \mid x]$

Regression function

- $y(x) = \mathbb{E}_t[t \mid x]$ – regression function
  - conditional average of t conditioned on x



- Recap:
  - Conditional expectation $\mathbb{E}_t[t \mid x]$
  - $\mathbb{E}_t[t \mid x] = \int t \frac{p(x,t)}{p(x)} \mathrm{dt} = \int t \frac{p(t \mid x)p(x)}{p(x)} \mathrm{dt} = \int t p(t \mid x) \mathrm{dt}$
  - **it does not depend on t !**

But why?

Why the optimal regression function is the conditional expectation, $\mathbb{E}_t[t\,|\,x]$?

- OK, let us derive the expected loss in slightly different way
- for fixed value of x:

$$(y(x)-t)^2 = (y(x)-\mathbb{E}[t\,|\,x]+\mathbb{E}[t\,|\,x]-t)^2 =$$

$$=\{y(x)-\mathbb{E}[t\,|\,x]\}^2 + 2\{y(x)-\mathbb{E}[t\,|\,x]\}\{\mathbb{E}[t\,|\,x]-t)\} + \{\mathbb{E}[t\,|\,x]-t)\}^2$$

- We put this into the formula below and integrate over t

$$\mathbb{E}[L] = \int\int (y(x)-t)^2 p(x,t)\mathrm{dx}\ \mathrm{dt}$$

- Hint: the cross-term should disappear

A FEW MOMENTS LATER

Great!

- $\mathbb{E}[L] = \int\int (y(x) - t)^2 p(x,t) \mathrm{dx}\ \mathrm{dt} =$

$$= \int \{y(x) - \mathbb{E}[t\,|\,x]\}^2 p(x)\ \mathrm{dx} + \int \{\mathbb{E}[t\,|\,x] - t)\}^2 p(x,t)\ \mathrm{dx}\ \mathrm{dt}$$

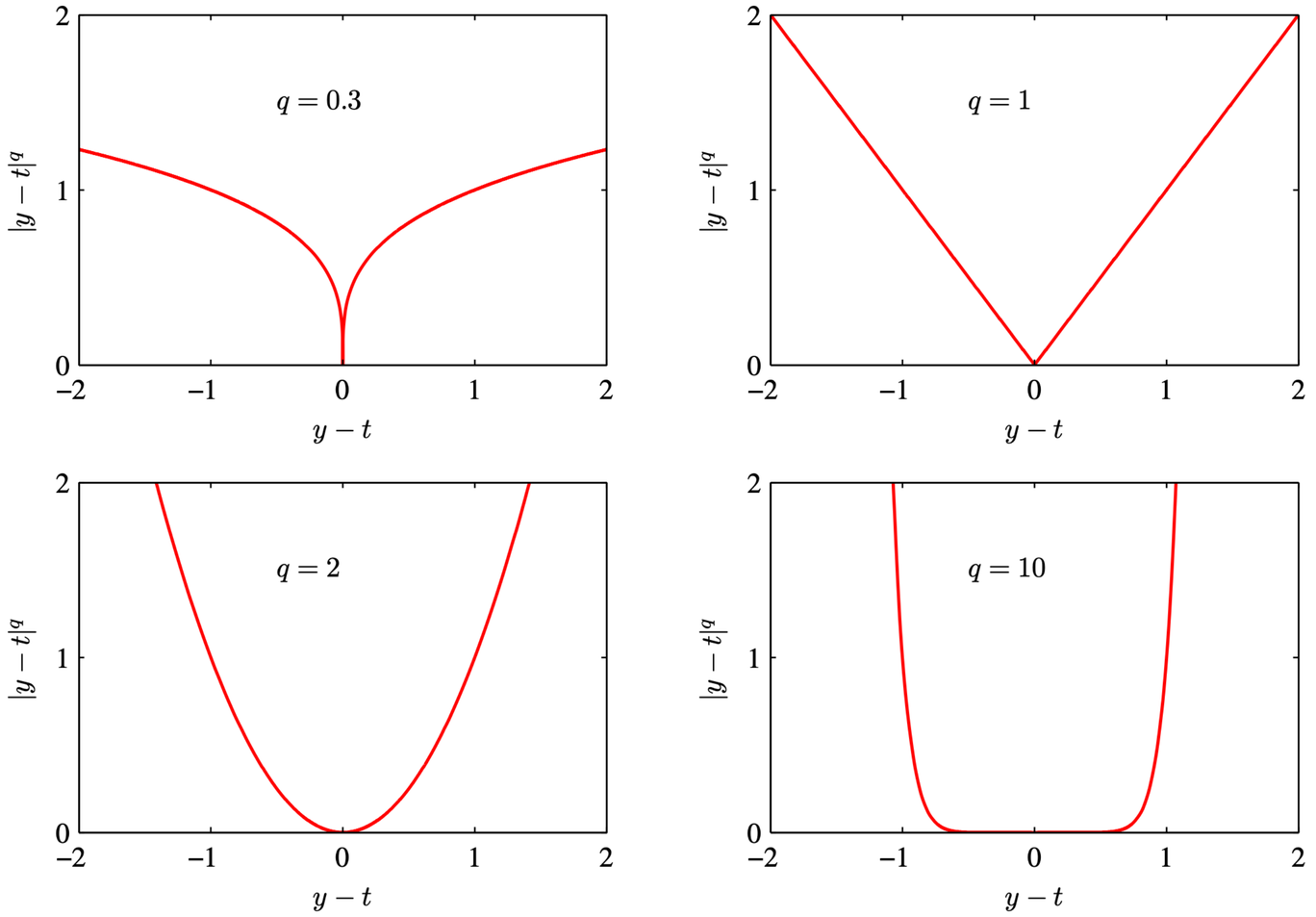
Therefore, the minimum of $\mathbb{E}[L]$ corresponds to $y(x) = \mathbb{E}_t[t\,|\,x]$ when the first term vanishes

The second term denotes an irreducible noise, or the variance of the distribution of target t, averaged over (all possible values of) x

Minkowski loss

- $\mathbb{E}[L_q] = \int\int |y(x) - t|^q p(x,t) \mathrm{dx}\ \mathrm{dt}$

- q=2: Conditional **Mean**
- q=1: Conditional **Median**
- $q \to \infty$: Conditional **Mode**



3 Approaches to Regression

1. First solve the inference problem of determining the joint density $p(x, t)$
2. First solve the inference problem of determining the conditional density $p(t|x)$
3. Find a regression function $y(x)$ directly from the training data

Break, 1 week

Summary

- Regression Models:
  - Loss function for regression
- Next week
  - Linear Regression
  - The Bias-Variance Decomposition

- Fill the feedback form with your questions, please

Statistical Techniques for Data Science & Robotics

Week 5

Quiz

- Recal and explain each term in the formula below

$$\mathbb{E}[L] = \int\int (y(x) - t)^2 p(x,t) \mathrm{dx}\ \mathrm{dt} =$$

$$= \int \{y(x) - \mathbb{E}[t\,|\,x]\}^2 p(x)\ \mathrm{dx} + \int\int \{\mathbb{E}[t\,|\,x] - t)\}^2 p(x,t)\ \mathrm{dx}\ \mathrm{dt}$$

Objectives

- Recap the Linear Regression
- Understand The Bias-Variance Decomposition
- Undestrand Fisher's Linear Discriminant

Linear regression

Setup

linear regression

- t is target
- **x** is input (D-dimensional vector)
- y(x) estimate of t for each x
- loss: L(t, y(x)) = $(y(x) - t)^2$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

In general (linear in w):

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Basis functions

a.k.a. features

- Polynomial:

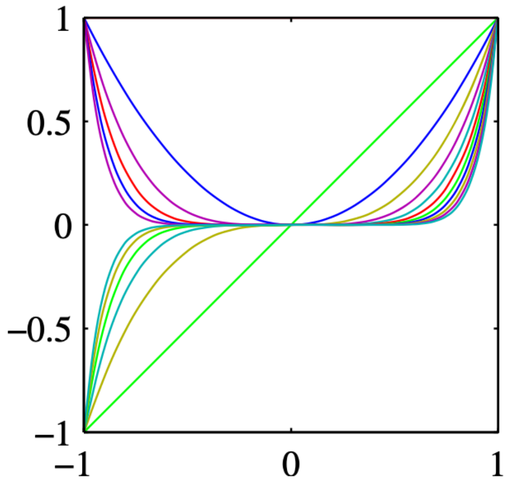
$$\phi_j(x) = x^j$$

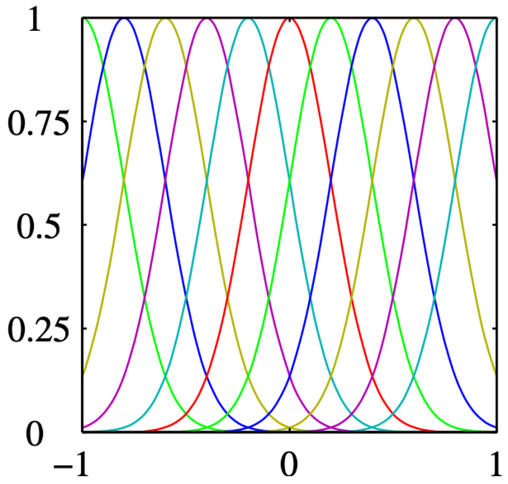
- Gaussian:

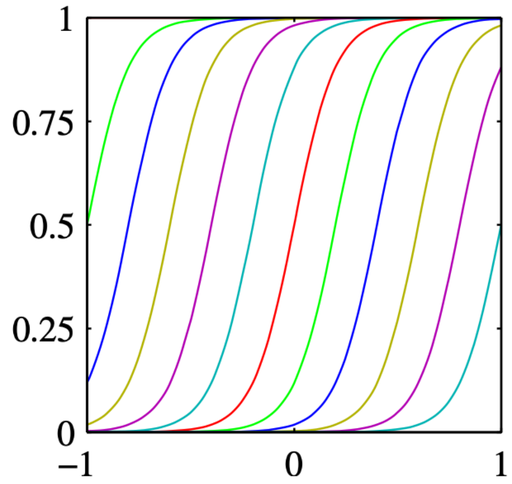
$$\phi_j(x) = \exp\{-\frac{(x-\mu_j)^2}{2s}\}$$

- Sigmoidal:

$$\phi_j(x) = \sigma\{\frac{x-\mu_j}{s}\}$$







$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Maximum likelihood and least squares

Build. Step by step

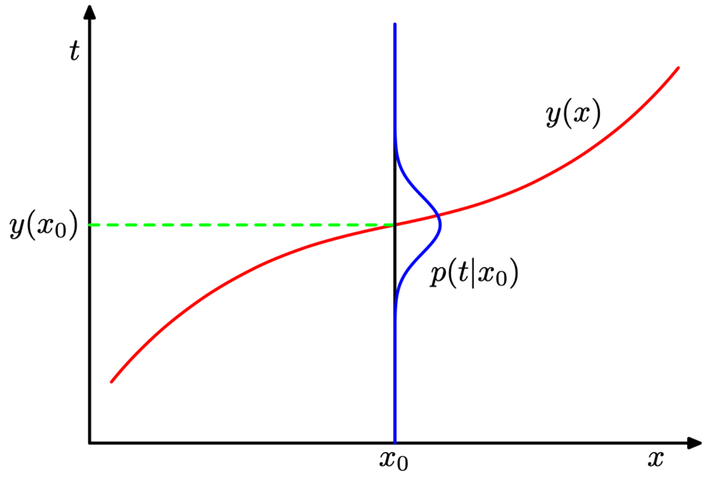
- t = y(x,w) + gaussian noise; (standard normal distr. with inverse variance $\beta = 1/\sigma^2$)
- Therefore, our model is $p(t \mid x, w, \beta) = \mathcal{N}(t \mid y(x, w), \beta^{-1})$
- likelihood function (N = sample size): $p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$
- log-likelihood:

$$\ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1})$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$

- Where

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$




A LITTLE LONGER
THAN A FEW
MINUTES LATER

Solution

- normal equations for the least squares problem

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

- $\mathbf{\Phi}$ is a design matrix (NxM)

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

Assumptions of simple linear regression

- Here we consider the **simplest** form: $$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

  - **Linearity of the data.** The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
  - **Normality of residuals.** The residual errors are assumed to be normally distributed.
  - **Homogeneity of residuals variance.** The residuals are assumed to have a constant variance (homoscedasticity)
  - **Independence** of residuals error terms.

http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/#regression-diagnostics-reg-diag

Potential problems

- **Non-linearity** of the outcome - predictor relationships
- **Heteroscedasticity:** Non-constant variance of error terms.
- **Presence of influential values** in the data that can be:
  - **Outliers**: extreme values in the outcome (y) variable
  - **High-leverage points:** extreme values in the predictors (x) variable

http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/#regression-diagnostics-reg-diag

The Bias-Variance Decomposition and Trade-off

Recall the Bias and Variance

For some estimator $\hat{\theta}$,

- $\hat{\theta}$ is an estimator of the parameter θ

- Bias:
  - $Bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$

- Variance:
  - $\mathbb{V}[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$

Setup

- t is target
- **x** is input (D-dim. vector)
- y(x) is estimate of t for each x
- Squared loss: $L(t, y(x)) = (y(x) - t)^2$

- For the squared loss function, the optimal prediction is given by the conditional expectation

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\, \mathrm{d}t.$$

We are minimizing the expected loss, $\mathbb{E}[L]$:

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

Goal

approximate the h(x) using the finite dataset,

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

- To minimize the first term we
  - need to find y(x)
  - be able to assess its quality,
    - e.g. how good is it on average over possible ensambles of datasets

Build

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

- Given a dataset, $\mathcal{D}$
- Our interest is this quantity, $\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2$
- which can be expanded: $\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2$
- $= \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2$
  $+2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.$
- Expectation $\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]$ is calculated via independent sampling of N datasets

Punchline

$$\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2 = \{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 + 2\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}.$$

- take the expectation of this expression with respect to D and note that the final term will vanish

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}$$

Expected loss decomposition

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

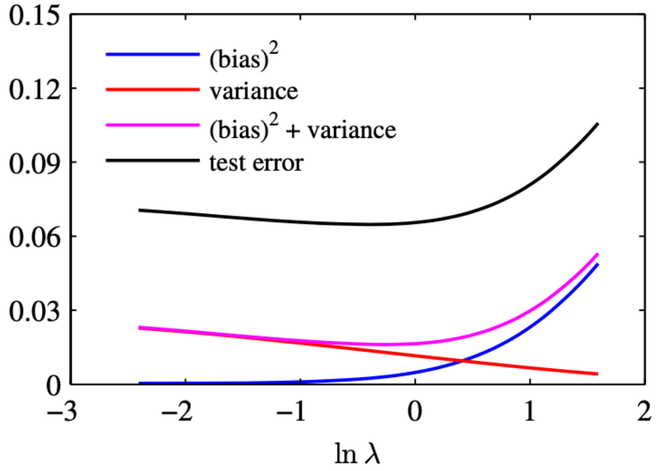
$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

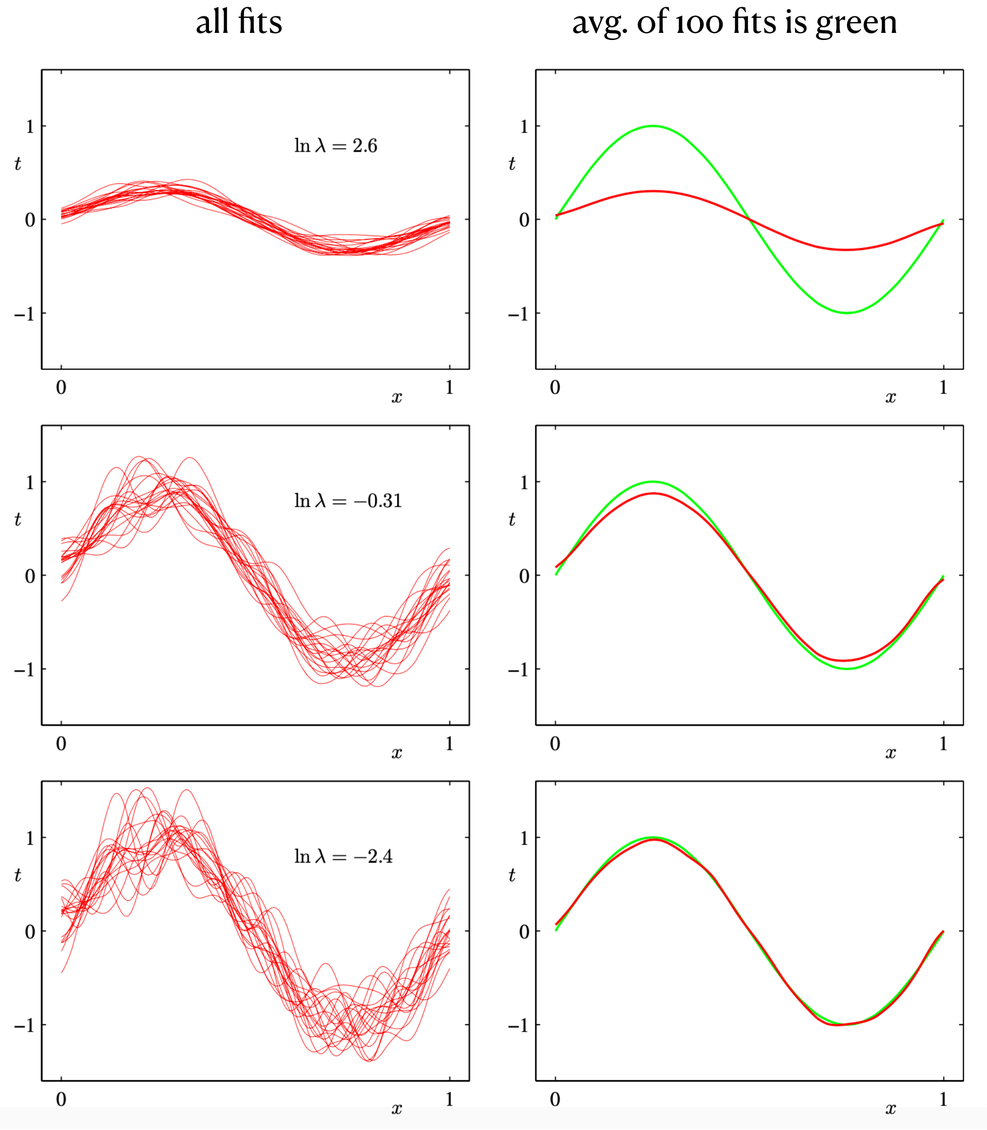
$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x},t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$$

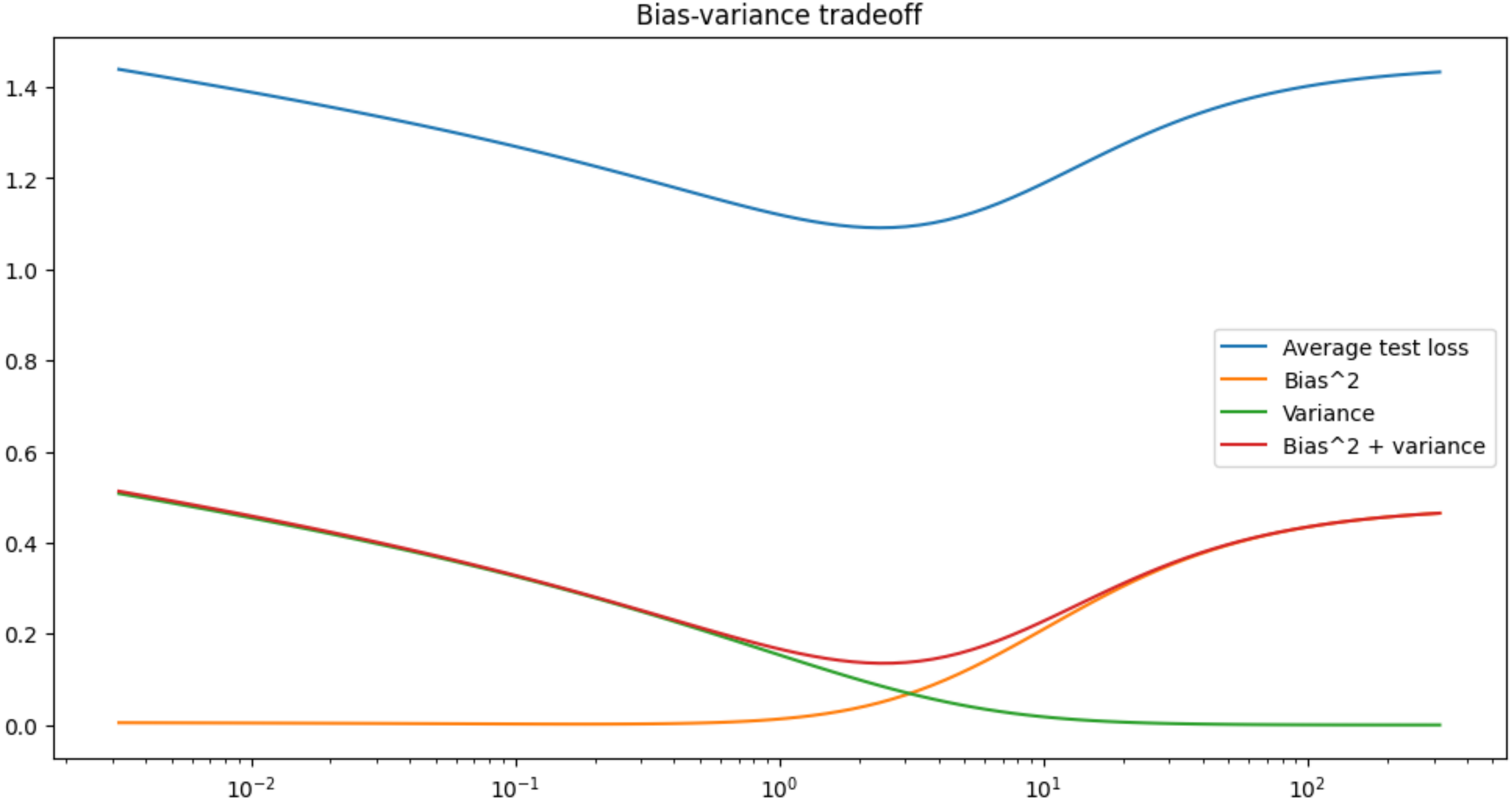
Example (Lab): Regularized least squares

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

- L = 100 datasets
- with N = 25 training examples each
- **Gaussian basis functions**
- λ is a regularizer coefficient
- Analysis:






Bias-variance tradeoff
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
Average test loss
Bias^2
Variance
Bias^2 + variance

Break, 5 min.

Linear models for Classification

Setup

- t is target, but $t \in \{0,1\}$
- **x** is input (D-dim. vector)
- y(x) is estimate of t for each x

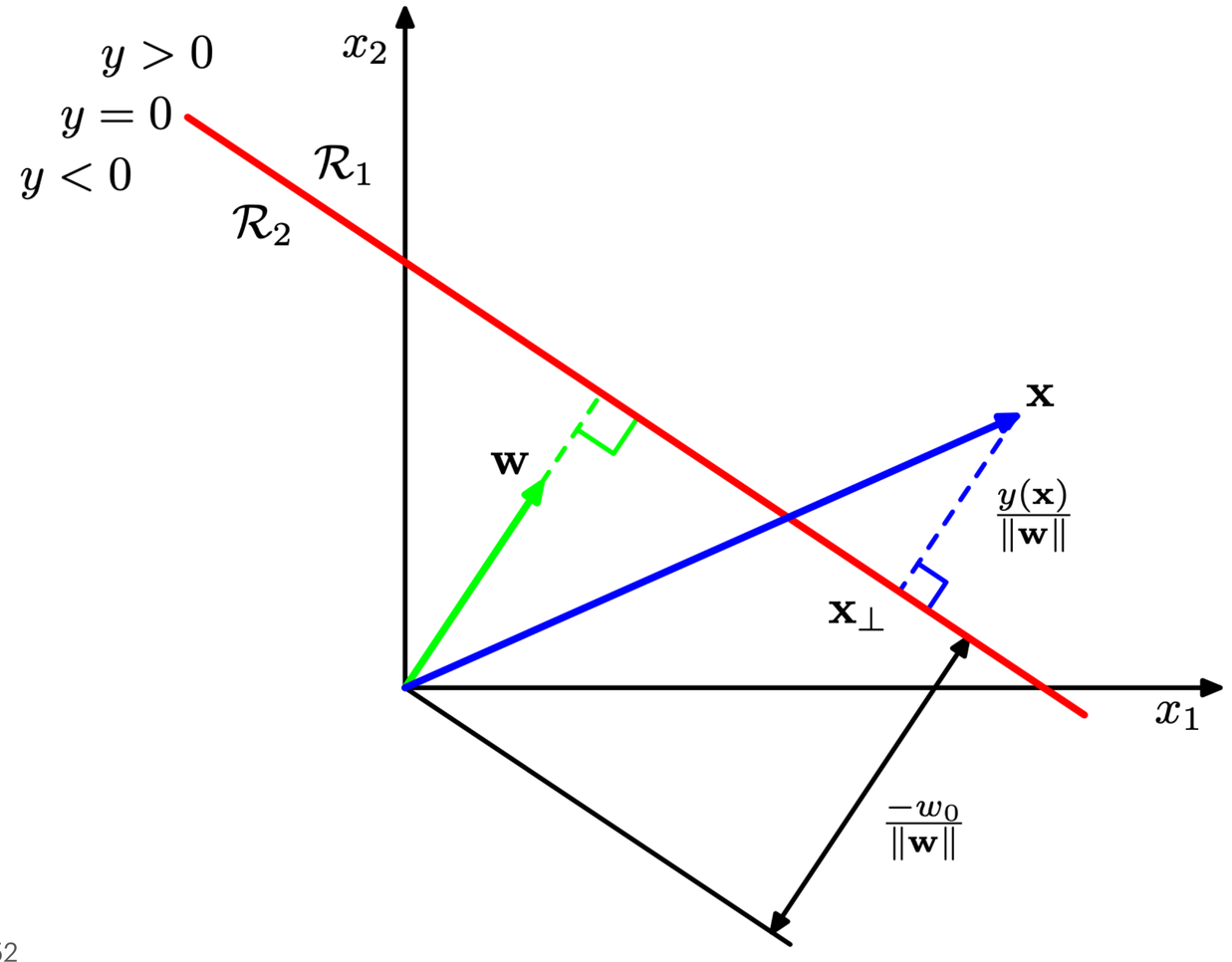
For example,

Generalized linear models

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0\right)$$

Discriminant Functions

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

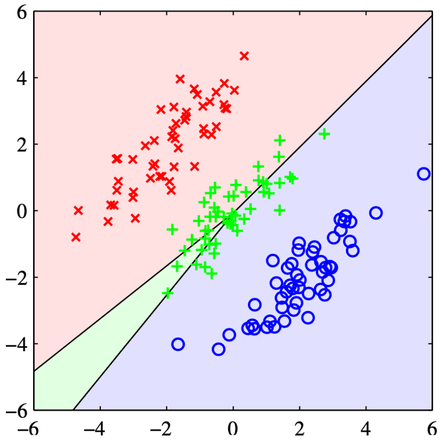


Least squares for classification

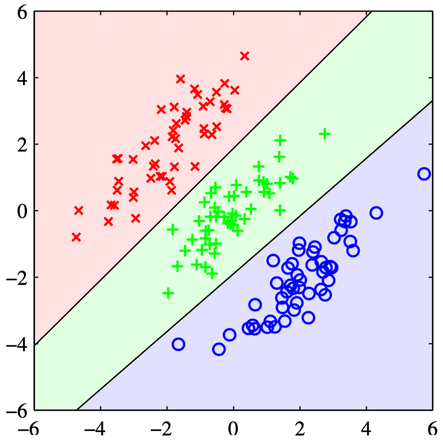
and problems

- for each class k
- $$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

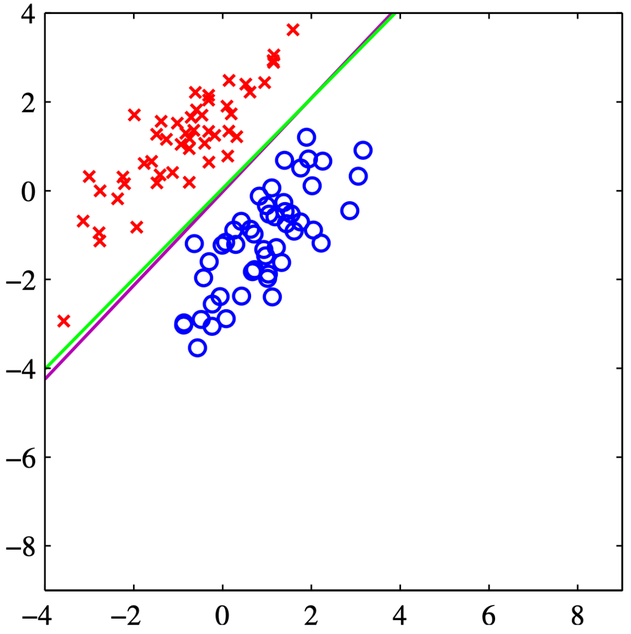
Logistic regression



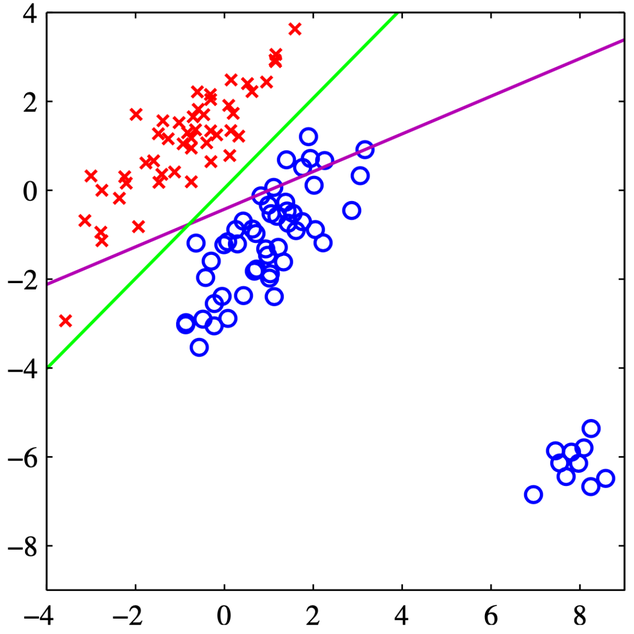
Least Squares



Logistic regression



Least Squares

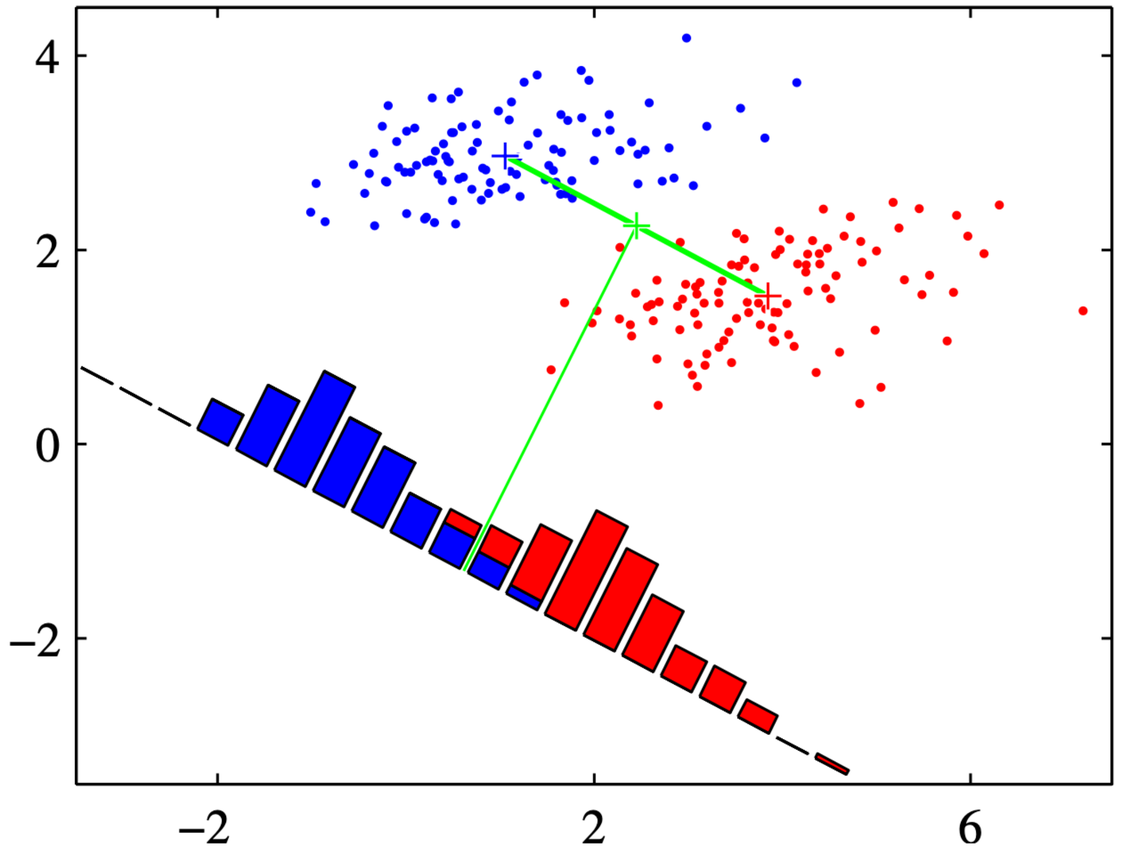


Fisher's linear discriminant

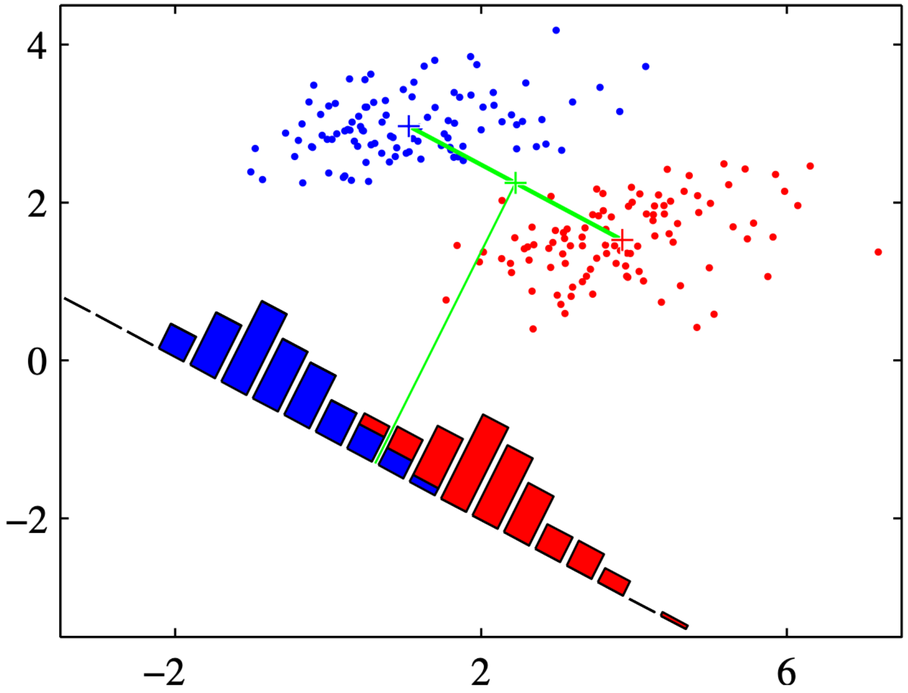
two classes, k = 2

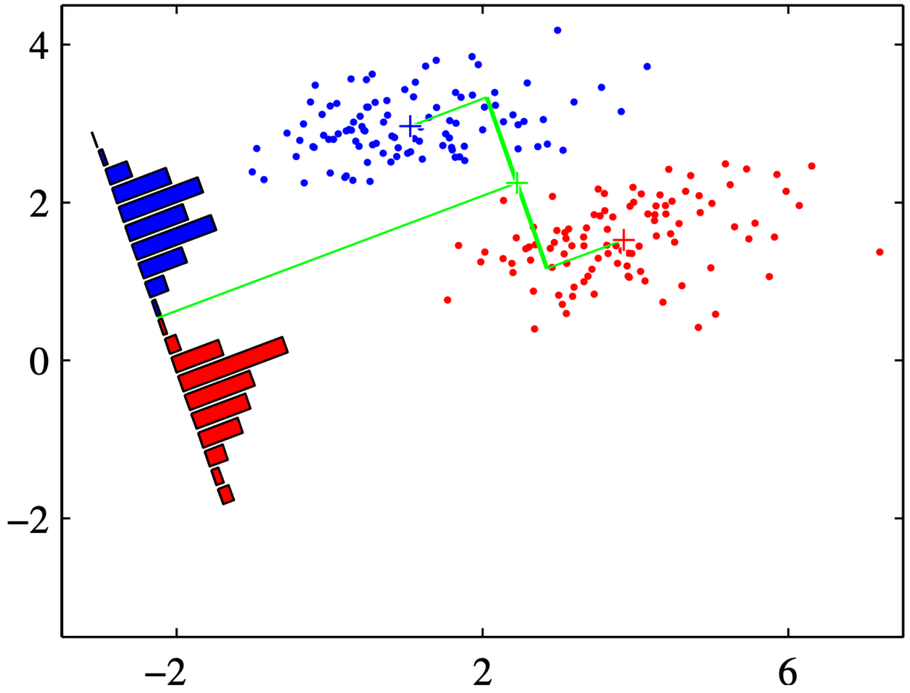
- Project high dimensional data to a line
  - and How to define that line?

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$



Which line is better?





Derivation of Fisher's Linear Discriminant

actually, only a direction of the line