MAP 569
Machine Learning II

Christophe Giraud

PC1

Supervised Classification

Supervised Classification

"Daily" supervised classification

1. **SPAM filter**
2. **Image recognition**: automatic postal ZIP reading
3. **Medical diagnosis**: cancers, alzheimer, etc
4. **In silico chemometrics**: research of some medicine
5. **Ad-online, recommandation, etc**

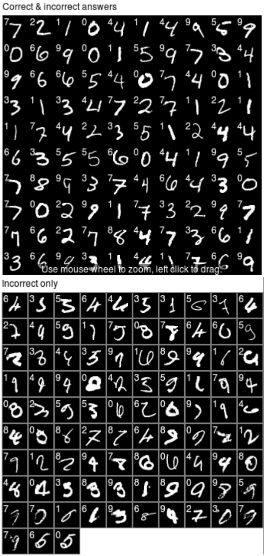
http://c-command.com/spamsieve/

"Daily" supervised classification

1. SPAM filter
2. **Image recognition**: automatic postal ZIP reading
3. **Medical diagnosis**: cancers, alzheimer, etc
4. **In silico chemometrics**: research of some medicine
5. **Ad-online, recommandation, etc**

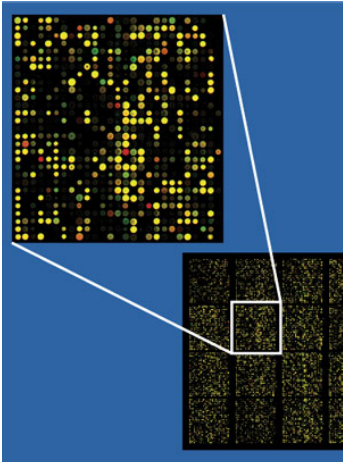
"Daily" supervised classification

1. **SPAM filter**
2. **Image recognition**: automatic postal ZIP reading
3. **Medical diagnosis**: cancers, alzheimer, etc
4. **In silico chemometrics**: research of some medicine
5. **Ad-online, recommandation, etc**



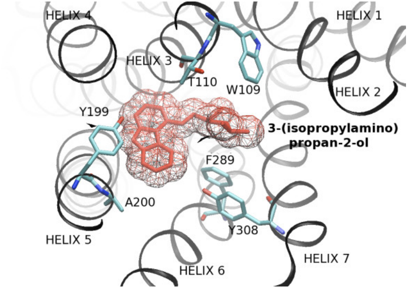
"Daily" supervised classification

1. **SPAM filter**
2. **Image recognition**: automatic postal ZIP reading
3. **Medical diagnosis**: cancers, alzheimer, etc
4. **In silico chemometrics**: research of some medicine
5. **Ad-online, recommandation, etc**



"Daily" supervised classification

1. **SPAM filter**
2. **Image recognition**: automatic postal ZIP reading
3. **Medical diagnosis**: cancers, alzheimer, etc
4. **In silico chemometrics**: research of some medicine
5. **Ad-online, recommandation, etc**

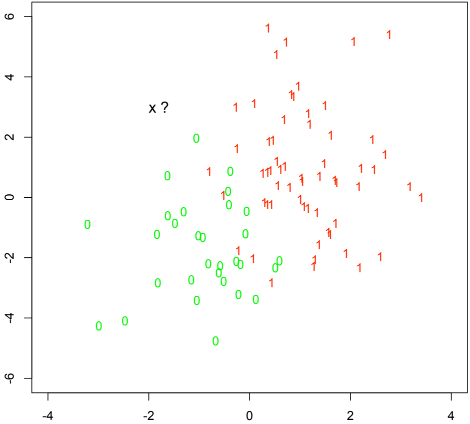


"Daily" supervised classification

1. **SPAM filter**
2. **Image recognition**: automatic postal ZIP reading
3. **Medical diagnosis**: cancers, alzheimer, etc
4. **In silico chemometrics**: research of some medicine
5. **Ad-online, recommandation, etc**

Framework

Observations: data points $X_i \in \mathcal{X}$ with labels $Y_i \in \{-1, 1\}$ for $i = 1, \ldots, n$.



Objective: predict the class of a new data point x.

Formalization

Classifier

Any (measurable) function $h : \mathcal{X} \to \{-1, 1\}$.

Risk

Probability of misclassification: $R(h) = \mathbb{P}(h(X) \neq Y)$

Bayes classifer

Check that the classifier $h_*(x) = \mathrm{sign}\left(\mathbb{P}\left[Y = 1|X = x\right] - 1/2\right)$ fulfills

$$R(h_*) = \min_h R(h).$$

Statistical issue

The distribution of (X, Y) is unknown. We only have an i.i.d. sample $(X_i, Y_i)_{i=1,\dots,n}$.

Parametric modeling

Modeling 1: parametric modeling of the distribution of (X, Y)

Example: Gaussian mixture

Model

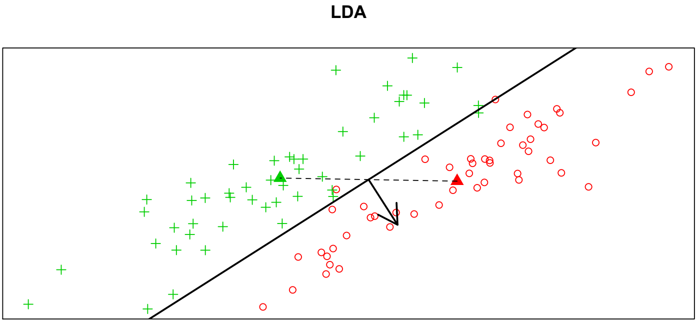
- $\mathbb{P}(Y_i = k) = \pi_k$, for $k = -1, 1$
- $\text{Distribution}(X_i | Y_i = k) \;=\; \mathcal{N}(\mu_k, \Sigma_k)$, for $k = -1, 1$.

Gaussian mixture

Exercise

1. What is the distribution of X?
2. Prove that the Bayes classifier is given by
$$h_*(x) = \mathrm{sign}\left(\pi_1 g_1(x) - \pi_{-1} g_{-1}(x)\right), \quad x \in \mathbb{R}^p.$$
3. Prove that when $\Sigma_{-1} = \Sigma_1 = \Sigma$, the condition $\pi_1 g_1(x) > \pi_{-1} g_{-1}(x)$ is equivalent to
$$(\mu_1 - \mu_{-1})^T \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2}\right) > \log(\pi_{-1}/\pi_1).$$
Interpret geometrically this result.

Gaussian mixture



Bayes Classifier

When $\Sigma_{-1} = \Sigma_1 = \Sigma$ we have

$$h_*(x) = 1 \iff \left(x - \frac{\mu_1 + \mu_{-1}}{2}\right)^T \Sigma^{-1}(\mu_1 - \mu_{-1}) > \log(\pi_{-1}/\pi_1).$$

In pratice: we estimate μ_{-1}, μ_1 and Σ by MLE

Mahalanobis distance

Exercise (continued)

1. If $\pi_1 = \pi_{-1}$, check that

$$\mathbb{P}(h_*(X) = 1 | Y = -1) = \Phi(-d(\mu_1, \mu_{-1})/2)$$

   where Φ is the cumulative distribution function of a standard Gaussian and $d(\mu_1, \mu_{-1})$ is the Mahalanobis distance defined by

$$d(\mu_1, \mu_{-1})^2 = (\mu_1 - \mu_{-1})^T \Sigma^{-1} (\mu_1 - \mu_{-1}).$$

2. When $\Sigma_1 \neq \Sigma_{-1}$, what is the nature of the frontier between $\{h_* = 1\}$ and $\{h_* = -1\}$?

Semi-parametric modeling

Bayes claissifier: $h_*(x) = \mathrm{sign}\left(\mathbb{P}\left[Y = 1|X = x\right] - 1/2\right)$

Modeling 2: modeling of the conditional distribution of Y given X

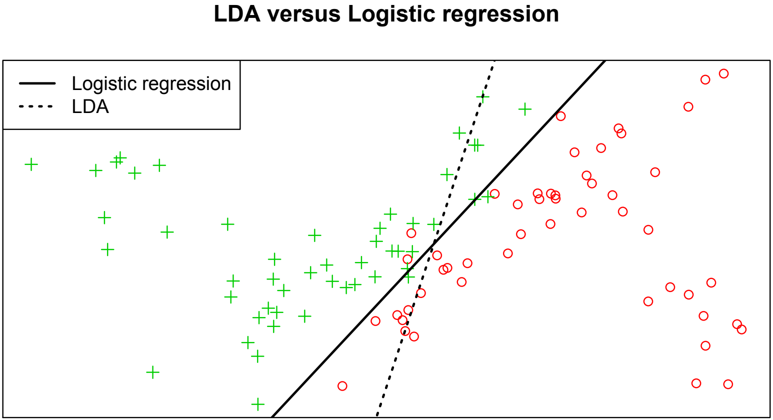
Logistic regression

$$\mathbb{P}\left[Y = 1|X = x\right] = \frac{\exp\left(\langle \beta^*, x\rangle\right)}{1 + \exp\left(\langle \beta^*, x\rangle\right)}$$

Bayes classifier

$$h_*(x) = 1 \iff \langle \beta^*, x\rangle > 0$$

Logistic regression



Maximum likelihood estimation

Conditional likelihood of Y given X

$$\widehat{\beta} \in \operatorname*{argmax}_{\beta \in \mathbb{R}^d} \prod_{i:Y_i=1} \left(\frac{\exp(\langle \beta, x_i \rangle)}{1 + \exp(\langle \beta, x_i \rangle)} \right) \prod_{i:Y_i=-1} \left(\frac{1}{1 + \exp(\langle \beta, x_i \rangle)} \right)$$

Logistic classifier

$\widehat{h}_{\text{logistic}}(x) = \operatorname{sign}\left(\langle \widehat{\beta}, x \rangle \right)$ for all $x \in \mathbb{R}^d$.

Synthetic data

- $\beta^* = (3, 0, -4, 0, 0.1)$
- x_{ij} i.i.d. standard Gaussian
- $n = 50$

```
> fit <- glm(y ~ pred1 + pred2 + pred3 + pred4 + pred5,
data=simulatedata, family=binomial())
> summary(fit)
```

	Estimate	Std. Error	z value	$Pr(>\lvert z\rvert)$	
pred1	3.3233	1.2205	2.723	0.00647	**
pred2	-0.6257	0.7885	-0.794	0.42745	
pred3	-4.7686	2.0019	-2.382	0.01722	*
pred4	-1.7596	1.1080	-1.588	0.11227	
pred5	-0.5450	0.7805	-0.698	0.48498	

Variable Selection

When $p \approx n$ or $p \gg n$,

1. we cannot trust asymptotics
2. we cannot implement model selection to select active variables

$$\Longrightarrow \widehat{\beta} \in \operatorname*{argmin}_{\beta} \left\{ \sum_{i=1}^{n} \ell \left(y_i (x_i^T \beta) \right) + \lambda |\beta|_1 \right\}$$