Week I

1. Notable Probability Distributions

·) Multinomial

Let's start from the simple case of a Bernoulli trial: You have 2 possible outcomes for an experiment/event, which can happen with probability π and $1-\pi$, respectively. You can think of the same experiment and record it as $(0,1)$ / $(1,0)$, i.e. the outcome is one of the two possible.

The multinomial distribution extends this to the case of multiple possible outcomes, i.e. the recorded outcome is one out of the K possible. For example, $K=4$

$(0,1,0,0)$, $(1,0,0,0)$, ..., $(0,0,1,0)$...

The binomial distribution counts the number of "successes" (1) in N independent Bernoulli trials.

The multinomial counts the number of successes in each category in N independent trials.

Let π_k be the probability of each trial resulting in category k. ($\sum_k \pi_k = 1$) Let $\underline{Y}$ be the vector of the number of successes in each category over N trials. Then,

$$\underline{Y} \sim \text{Multinomial}(N; \pi_1, \dots, \pi_K) \text{ and}$$

the probability density function is

Pdf of Multinomial dist →

$$P(\underline{Y} \mid N, \underline{\pi}) = \frac{N!}{Y_1! Y_2! \dots Y_K!} \pi_1^{Y_1} \pi_2^{Y_2} \dots \pi_K^{Y_K}$$

NOTE THAT THE MARGINAL $Y_k \sim \text{Binomial}(N, \pi_k)$

AND $\mathbb{E}[\underline{Y}] = N\underline{\pi}$

$\text{Cov}(Y_i, Y_j) = -N\pi_i\pi_j \quad , i \neq j$

•) MULTIVARIATE NORMAL (GAUSSIAN)

This is a continuous probability distribution.

p = dimension of $\underline{Y}$

$$\underline{Y} \sim N_p(\underline{\mu}, \Sigma)$$

$\underline{\mu}$: MEAN

Σ: COVARIANCE MATRIX: For any linear combination (non-random) $\underline{c}$,

$$\text{Var}(\underline{c}^T\underline{Y}) = \text{Var}(c_1Y_1 + \cdots + c_pY_p) = \underline{c}^T\Sigma\underline{c}$$

The probability density function is

$$p(\underline{Y} \mid \underline{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{Y}-\underline{\mu})^T\Sigma^{-1}(\underline{Y}-\underline{\mu})}$$

REMARK: Σ^{-1} is often called PRECISION MATRIX.

- The marginal distributions are

$$Y_j \sim N(\mu_j, \sigma_j^2), \quad \text{where } \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & & \\ \sigma_{21} & \ddots & & \\ & & & \sigma_p^2 \end{bmatrix}$$

AND $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$

- Any LINEAR COMBINATION HAS NORMAL DISTRIBUTION:

$$\underline{c}^T\underline{Y} \sim N(\underline{c}^T\underline{\mu}, \underline{c}^T\Sigma\underline{c})$$

→ Partitioned Gaussian

Let's now consider a partition $\underline{Y} = \begin{bmatrix} \underline{Y}_1 \\ \underline{Y}_2 \end{bmatrix} \begin{matrix} \} p_1 \\ \} p_2 \end{matrix}$ ⇔ dimensions

$$\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \qquad (\Sigma_{12} = \Sigma_{21}^T)$$

and $\Sigma^{-1} = \Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$.

- **The marginal density of $\underline{Y}_1$:**

$$\left[p(\underline{Y}_1) = \int p(Y_1, Y_2)\, dY_2 \right]$$

Let $\delta_{\underline{J}} = \underline{Y} - \underline{\mu} = \begin{bmatrix} \delta_{\underline{J}_1} \\ \delta_{\underline{J}_2} \end{bmatrix}$

Computing Sq. forms on this relation using ⊛ next pg.

Now, $-\frac{1}{2} \delta_J^T \Lambda \delta_J = -\frac{1}{2} \begin{bmatrix} \delta_{J_1}^T & \delta_{J_2}^T \end{bmatrix} \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} \delta_{J_1} \\ \delta_{J_2} \end{bmatrix} =$

$$= -\frac{1}{2} \delta_{\underline{J}_1}^T \Lambda_{11} \delta_{\underline{J}_1} - \frac{1}{2} \delta_{\underline{J}_2}^T \Lambda_{22} \delta_{\underline{J}_2} - \delta_{J_1}^T \Lambda_{12} \delta_{J_2} =$$

$$(*) = -\frac{1}{2} \delta_{\underline{J}_1}^T \Lambda_{11} \delta_{\underline{J}_1} - \frac{1}{2} \left(\delta_{J_2} + \Lambda_{22}^{-1} \Lambda_{21} \delta_{J_1} \right)^T \Lambda_{22} \left(\delta_{J_2} + \Lambda_{22}^{-1} \Lambda_{21} \delta_{J_1} \right)$$

$$+ \frac{1}{2} \delta_{J_1}^T \Lambda_{12} \Lambda_{22}^{-1} \Lambda_{21} \delta_{\underline{J}_1}$$

(2, Sg.)

I WANT TO WRITE THE SECOND PART AS A QUADRATIC FORM IN $\delta_{\underline{J}_2}$

PLUS A TERM THAT DEPENDS ONLY ON $\delta_{\underline{J}_1}$

(BECAUSE I AM GOING TO INTEGRATE W.R.T. $\delta_{\underline{J}_2}$)

Aside: Completing the Square

For scalars: $\frac{a}{2}x^2 + bx = \frac{a}{2}\left(x^2 + \frac{2b}{a}x\right) =$

$$= \frac{a}{2}\left[\left(x + \frac{b}{a}\right)^2 - \frac{b^2}{a^2}\right]$$

For vectors: (A symmetric)

$$\frac{1}{2}\underline{x}^T A \underline{x} + \underline{b}^T \underline{x} = \frac{1}{2}(\underline{x} + \underline{m})^T A (\underline{x} + \underline{m}) - c =$$

$$= \frac{1}{2}\underline{x}^T A \underline{x} + \frac{1}{2}\underline{x}^T A \underline{m} + \frac{1}{2}\underline{m}^T A \underline{x} + \frac{1}{2}\underline{m}^T A \underline{m} - c =$$

$$= \frac{1}{2}\underline{x}^T A \underline{x} + \underline{m}^T A \underline{x} + \frac{1}{2}\underline{m}^T A \underline{m} - c$$

$\frac{1}{2}\underline{m}^T A \underline{m} - c$ set $= 0$

$$\Rightarrow \underline{b}^T = \underline{m}^T A \Rightarrow \underline{m} = A^{-1}\underline{b}$$

$$c = \frac{1}{2}\underline{m}^T A \underline{m} = \frac{1}{2}\underline{b} A^{-1} \underline{b}$$

(*) $\Rightarrow$

$$\frac{1}{2}\underline{x}^T A \underline{x} + \underline{b}^T \underline{x} = \frac{1}{2}\left(\underline{x} + A^{-1}\underline{b}\right)^T A \left(\underline{x} + A^{-1}\underline{b}\right) + - \frac{1}{2}\underline{b}^T A^{-1}\underline{b}$$

↑ Completing Square formula for vectors

Now let's integrate: ← ignore first term of PDF of Normal & focus on exponent term only.

$$\int p(\underline{y}_1, \underline{y}_2)\, d\underline{y}_2 \propto \int \exp\left(-\frac{1}{2}\underline{\delta}_{\underline{y}}^{\top} \Lambda\, \underline{\delta}_{\underline{y}}\right) d\underline{y}_2 =$$

$$= \exp\left(-\frac{1}{2}\underline{\delta}_{\underline{y}_1}^{\top} \Lambda_{11}\, \underline{\delta}_{\underline{y}_1} + \frac{1}{2}\underline{\delta}_{\underline{y}_1}^{\top} \Lambda_{12} \Lambda_{22}^{-1} \Lambda_{21}\, \underline{\delta}_{\underline{y}_1}\right) \cdot$$

Term depending on $\underline{y}_1$ comes out of integral ↙

$$\int \exp\left[-\frac{1}{2}\left(\underline{\delta}_{\underline{y}_2} + \Lambda_{22}^{-1}\Lambda_{21}\underline{\delta}_{\underline{y}_1}\right)^{\top} \Lambda_{22}\left(\underline{\delta}_{\underline{y}_2} + \Lambda_{22}^{-1}\Lambda_{21}\underline{\delta}_{\underline{y}_1}\right)\right] d\underline{y}_2$$

↑ This integral gives 1 but constants come out as

$$\propto \text{const} \cdot N_{p_2}\left(\underline{\mu}_2 - \Lambda_{22}^{-1}\Lambda_{21}\underline{\delta}_{\underline{y}_1},\ \Lambda_{22}^{-1}\right)$$

$$\Rightarrow \int p(\underline{y}_1, \underline{y}_2)\, d\underline{y}_2 \propto (2\pi)^{\frac{p_2}{2}} \left|\Lambda_{22}^{-1}\right|^{\frac{1}{2}} \exp\left(-\frac{1}{2}\underline{\delta}_{\underline{y}_1}\Lambda_{11}\underline{\delta}_{\underline{y}_1} + \frac{1}{2}\underline{\delta}_{\underline{y}_1}^{\top}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\underline{\delta}_{\underline{y}_1}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\underline{y}_1 - \underline{\mu}_1\right)\left(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\right)\left(\underline{y}_1 - \underline{\mu}_1\right)\right)$$

$$\Rightarrow \underline{Y}_1 \sim N\left(\underline{\mu}_1, \left(\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}\right)^{-1}\right)$$

↓

Marginal density of $\underline{Y}_1$

→ Written also in most Papers using Block inverses.

Slide 2 : Block inverse

Block matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$, A, D square

$$\begin{bmatrix} M & N \\ P & Q \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = ?$$

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} M & N \\ P & Q \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

$\Rightarrow$ $AM + BP = I$, $CM + DP = 0$, $\Rightarrow DP = -CM$

$$P = -D^{-1}CM$$

$$AM - BD^{-1}CM = I \Rightarrow M = (A - BD^{-1}C)^{-1}$$

and $$P = -(D - CA^{-1}B)^{-1}CA^{-1}$$

In our case: $\Lambda^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{12} & \Lambda_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{12} \end{bmatrix}$

(A, B, C, D on the left; M, N, P, Q on the right)

(M) $$\Sigma_{11} = (\Lambda_{11} - \Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21})^{-1}$$

$\Rightarrow$ $\underline{Y}_1 \sim N_{p_1}(\underline{M}_1, \Sigma_{11})$

Marginal density of $\underline{Y}_1$

Main idea

- Marginal dist. & Conditional dist of Multivariate Normal

End of Week I

Week II

Conditional Density of Multivariate Normal (from last week)

$$p(\underline{y}_2 \mid \underline{y}_1) = \frac{p(y_1, y_2)}{p(y_1)}$$

Now, $\frac{p(y_1, y_2)}{p(y_1)} \propto p(y_1, y_2)$ w.r.t. y_2

$$\propto \exp\left(-\frac{1}{2}\, \delta_{\underline{y}_1}' \Lambda\, \delta_{\underline{y}_1}\right) \qquad \delta_{y_1} = y_1 - \mu_1$$

$$\Rightarrow p(\underline{y}_2 \mid \underline{y}_1) \propto \exp\left(-\frac{1}{2}\left(\delta_{\underline{y}_2} + \Lambda_{22}^{-1}\Lambda_{21}\delta_{\underline{y}_1}\right)' \Lambda_{22}\left(\delta_{\underline{y}_2} + \Lambda_{22}^{-1}\Lambda_{21}\delta_{\underline{y}_1}\right)\right)$$

$$= \exp\left[-\frac{1}{2}\left(\underline{y}_2 - \underline{\mu}_{2|1}\right)' \Lambda_{22}\left(\underline{y}_2 - \underline{\mu}_{2|1}\right)\right]$$

where $\mu_{2|1} = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\underline{y}_1 - \underline{\mu}_1\right)$

From Block inverse, but:

1) $\Lambda_{22} = \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}$

2) $\Lambda_{21} = -\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1}$

$$\Rightarrow \underline{\mu}_{2|1} = \mu_2 - \underbrace{\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)}_{\Lambda_{22}^{-1}}\underbrace{\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}}\Sigma_{21}\Sigma_{11}^{-1}\left(\underline{y}_1 - \underline{\mu}_1\right) = -\Lambda_{21}$$

(Two red terms give identity as their inverses)

$$= \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}\left(\underline{y}_1 - \underline{\mu}_1\right)$$

Conditional density

$$\Rightarrow \underline{y}_2 \mid \underline{y}_1 = y_1 \sim N_{p_2}\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}\left(y_1 - \mu_1\right),\ \underbrace{\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}}_{\Lambda_{22}^{-1}}\right)$$

2 ESTIMATION

- MLE → Frequentist
- Bayesian estimation for Gaussian

TWO PARADIGMS < FREQUENTIST / BAYESIAN

FREQUENTIST

prob. dist

$$Y \sim p(Y;\theta)$$

θ UNKNOWN PARAMETER

observations

SAMPLES: $Y_1, \dots, Y_N \sim Y$

ESTIMATOR: FUNCTION OF THE SAMPLE $\hat{\theta} = g(Y_1, \dots, Y_n)$

(which hopefully is telling us something about θ)

We choose the estimator based on some (frequentist) properties:

- SMALL MSE $= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$ ← MSE = Mean Sq. Error
- $\hat{\theta} \to \theta$ when $N \to \infty$

However, there are criterions to judge an estimator but they do NOT tell us how to pick one.

There are a few CONSTRUCTIVE procedures to build estimators with good (heuristic) properties, one of which is MAXIMUM LIKELIHOOD ESTIMATION

MLE in general:

$$\hat{\theta} = \arg\max L(\theta; Y_1, \dots, Y_N) \overset{\text{INDEPENDENT SAMPLES}}{=} \arg\max \prod_{i=1}^{N} p(Y_i;\theta)$$

MLE FOR GAUSSIAN SAMPLES

- Likelihood = f'n of parameters given sample
- Pdf = f'n of variables having certain parameters

i.i.d. = INDEPENDENT AND IDENTICALLY DISTRIBUTED

$$\underline{Y}_i \overset{i.i.d.}{\sim} N_p(\underline{\mu}, \Sigma), \quad i = 1, \dots, N$$

$$\vartheta = (\mu, \Sigma)$$

Gaussian likelihood

$$L(\mu, \Sigma; \underline{Y}) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\underline{Y}_i - \underline{\mu})^T \Sigma^{-1} (\underline{Y}_i - \underline{\mu})\right)$$

$$= \frac{1}{(2\pi)^{Np/2} |\Sigma|^{N/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (\underline{Y}_i - \underline{\mu})^T \Sigma^{-1} (\underline{Y}_i - \underline{\mu})\right)$$

To maximize it, it is useful to apply a log-transform:

$$\ell = \log(L) = -\frac{Np}{2}\log(2\pi) - \frac{N}{2}\log(|\Sigma|) - \frac{1}{2}\sum_{i=1}^{N} (\underline{Y}_i - \underline{\mu})^T \Sigma^{-1} (\underline{Y}_i - \underline{\mu})$$

Now, $(\underline{Y}_i - \underline{\mu})^T \Sigma^{-1} (\underline{Y}_i - \underline{\mu}) = \underline{Y}_i^T \Sigma^{-1} \underline{Y}_i - 2\underline{Y}_i^T \Sigma^{-1} \underline{\mu} + \underline{\mu}^T \Sigma^{-1} \underline{\mu}$

To find the maximum: differentiate log-likelihood & set to zero.

$$\frac{\partial}{\partial \underline{\mu}} \ell = 0 \Rightarrow$$

$$-\frac{1}{2}\sum_{i=1}^{N} \left(-2\Sigma^{-1}\underline{Y}_i + 2\Sigma^{-1}\underline{\mu}\right) = 0$$

$$\Sigma^{-1}\left(\sum_{i=1}^{N} \underline{Y}_i - N\underline{\mu}\right) = 0$$

$$\hat{\underline{\mu}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} \underline{Y}_i$$

NOTE: IT IS A MAXIMUM BECAUSE ℓ IS A POSITIVE DEFINITE QUADRATIC FORM IN μ.

for Σ, $\frac{\partial}{\partial \Sigma} \ell = 0 \Rightarrow$ $\hat{\Sigma}_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} (\underline{Y}_i - \underline{\mu})^T (\underline{Y}_i - \underline{\mu})$ //

(left for exercise)

ALSO NOTE: $\hat{\underline{\mu}} \sim N_p(\underline{\mu}, \frac{1}{N}\Sigma)$ (PROPERTY OF GAUSSIAN DISTRIBUTION)

BAYESIAN ESTIMATION:

You have prior knowledge, or belief, about the parameter θ that can be modelled with some probability distribution.

$\theta \sim p(\theta)$

Then, you can use Bayes theorem to update your ~~belief~~ belief about θ:

$$(*) \quad \underbrace{p(\theta \mid Y_1, \dots, Y_N)}_{\text{POSTERIOR}} \propto \underbrace{p(Y_1, \dots, Y_N \mid \theta)}_{\text{LIKELIHOOD}} \underbrace{p(\theta)}_{\text{PRIOR}}$$

and you get a posterior distribution for θ.

BAYESIAN ESTIMATION FOR GAUSSIAN DISTRIBUTIONS

- Σ KNOWN (FOR SIMPLICITY):

PRIOR: $\underline{\mu} \sim N_p(\underline{\mu}_0, \Sigma_0)$

POSTERIOR: $p(\underline{\mu} \mid \underline{Y}) \propto p(\underline{Y} \mid \mu) p(\underline{\mu}) = \dots$

↑ using * above

$$\propto \exp\left[-\frac{1}{2}\sum_{i=1}^{N}(\underline{y}_i-\underline{\mu})^T\Sigma^{-1}(\underline{y}_i-\underline{\mu})\right]\exp\left[-\frac{1}{2}(\underline{\mu}-\underline{\mu}_0)^T\Sigma_0^{-1}(\underline{\mu}-\underline{\mu}_0)\right]$$

$$= \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\underline{y}_i^T\Sigma^{-1}\underline{y}_i - 2\sum_{i=1}^{N}\underline{y}_i^T\Sigma^{-1}\underline{\mu} + N\underline{\mu}^T\Sigma^{-1}\underline{\mu} + \underline{\mu}^T\Sigma_0^{-1}\underline{\mu} + \underline{\mu}_0^T\Sigma_0^{-1}\underline{\mu}_0 - 2\underline{\mu}_0^T\Sigma_0^{-1}\underline{\mu}\right)\right]$$

(IGNORING TERMS DOES NOT DEPEND ON $\underline{\mu}$)

$$\propto \exp\left(-\frac{1}{2}\left[\underline{\mu}^T\left(N\Sigma^{-1}+\Sigma_0^{-1}\right)\underline{\mu} - 2\left(\sum_{i=1}^{N}\underline{y}_i^T\Sigma^{-1}\underline{\mu} + \underline{\mu}_0^T\Sigma_0^{-1}\underline{\mu}\right)\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\left(\underline{\mu}-\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\underline{\mu}_0+N\Sigma^{-1}\hat{\underline{\mu}}_{MLE}\right)\right)^T\cdot\right.$$

$$\left.\cdot\left(N\Sigma^{-1}+\Sigma_0^{-1}\right)\left(\underline{\mu}-\left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\underline{\mu}_0+N\Sigma^{-1}\hat{\underline{\mu}}_{MLE}\right)\right)\right]$$

Bayesian Estimation of μ for Gaussian

$$\Rightarrow \begin{cases} \underline{\mu}\,|\,\underline{y} \sim N_p\left(\underline{\mu}_N, \Sigma_N\right) \\ \underline{\mu}_N = \left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1}\left(\Sigma_0^{-1}\underline{\mu}_0+N\Sigma^{-1}\hat{\underline{\mu}}_{MLE}\right) \\ \Sigma_N = \left(\Sigma_0^{-1}+N\Sigma^{-1}\right)^{-1} \end{cases}$$

IF YOU WANT A POINT ESTIMATOR (AS OPPOSED TO A DISTRIBUTION), $\underline{\mu}_N$ IS THE NATURAL CHOICE (HERE IT IS BOTH THE MEAN AND THE MODE OF THE POSTERIOR) MORE MODE

NOTE ALSO:

$\underline{\mu}_N \to \hat{\underline{\mu}}_{MLE}$ IF $N \to +\infty$ ← more data & less relative importance of prior

$\underline{\mu}_N \to \underline{\mu}_0$ prior IF $N \to 0$ ← less data & more relative importance of prior

- GAUSSIAN PRIORS ARE CALLED CONJUGATE PRIOR FOR THE GAUSSIAN LIKELIHOOD (POSTERIOR CAN BE OBTAINED WITH SIMPLE UPDATE RULE)

- IF Σ IS UNKNOWN,

PRIOR : $\underline{M} \sim N_1\left(M_0, \frac{1}{\lambda}\Sigma\right)$

$\Sigma \sim$ inverse wishart $W^{-1}(\Psi, \nu)$

POSTERIOR : $\underline{M} \mid \underline{Y} \sim N\left(\tilde{M}, \frac{1}{\lambda}\Sigma\right)$

$$\Sigma \mid \underline{Y} \sim W^{-1}\left(\hat{\Sigma} + \Psi, N + \nu\right)$$

$$\tilde{M} = \frac{\lambda \underline{M}_0 + N \hat{M}_{MLE}}{\lambda + N}$$

- IN PRACTICE, CONJUGATE PRIOR ARE NOT USED ANYMORE.

COMPUTATIONAL METHODS ARE USED TO SAMPLE FROM POSTERIOR DISTRIBUTIONS;

↳ MARKOV CHAIN MONTE CARLO (MCMC)

FOR WHICH WE DO NOT HAVE EXPLICIT EXPRESSION.

- If I have more data & less certainty of prior then M^* will be the MLE.
& If I have less of data & more certainty of Prior then M^* will be closer to my prior.

Decision Theory

One way to choose estimators from posterior distributions is with decision theory:

Let $\tilde{\mu}$ be an estimator of μ and assume a quadratic loss, i.e.

$$L(\tilde{\mu}_i) = (\hat{\mu}_i - \mu_i)^2 \quad \text{for } i = 1, \dots, p$$

We would like to minimize L, but μ is unknown. So we minimize the posterior expected loss (also called Bayes Risk)

$$\Rightarrow \mathbb{E}_{\mu|y}[L(\hat{\mu}_3)] = \mathbb{E}_{\mu_3|y}[L(\tilde{\mu}_3)] = \mathbb{E}_{\mu_3|y}[(\hat{\mu}_3 - \mu_3)^2] =$$

$$= \mathbb{E}_{\mu_3|y}[((\tilde{\mu}_3 - \mu_{*3}) + (\mu_{*3} - \mu_3))^2] =$$

← add & sub max posterior μ^* as before from Bayesian estimation.

$$= \mathbb{E}_{\mu_3|y}[(\tilde{\mu}_3 - \mu_{*3})^2] + (\mu_{*3} - \hat{\mu}_3)^2 + 2(\mu_{*3} - \hat{\mu}_3)\mathbb{E}[\mu_3 - \mu_{*3}]$$

$$= V_{\mu_3|y}(\mu_3) + (\mu_{*3} - \hat{\mu}_3)^2$$

0, because $\mathbb{E}[\mu_*] = \mu$

This is minimized for $\hat{\mu}_3 = \mu_{*3}$

↑ minimized

↑ Variance of posterior μ_3

End of Week II

3. Gaussian Mixture Models

Week III

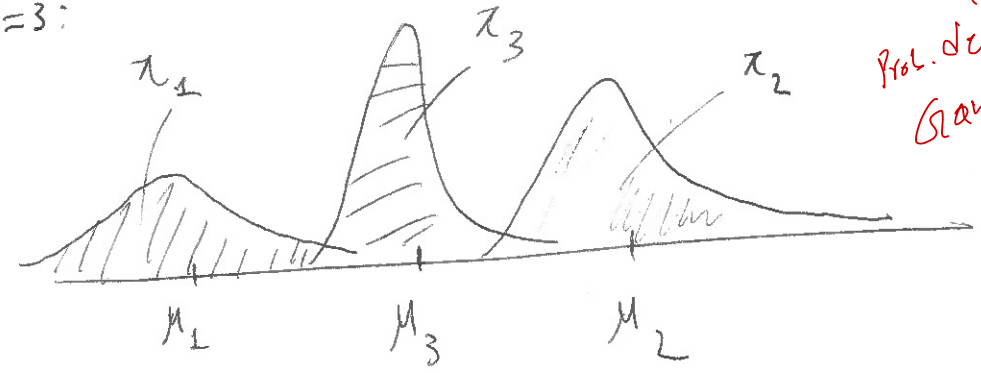
↳ & Parameter estimation

Random variables that are "mixture of normals", i.e. they have probability distribution

$$p(\underline{y};\theta) = \sum_{k=1}^{K} \pi_k \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\underline{y}-\underline{\mu}_k)^T \Sigma_k^{-1} (\underline{y}-\underline{\mu}_k)\right)$$

↑ Prob. density of Gaussian mixture

In 1D; K=3:



$$\sum_{k=1}^{K} \pi_k = 1, \quad \underline{\theta} = (\pi_1, \ldots, \pi_K, \underline{\mu}_1, \ldots, \underline{\mu}_K, \Sigma_1, \ldots, \Sigma_K)$$

Alternative representation:

Let $\underline{w}$ be a random vector such that:

$$w_k = \begin{cases} 1 & \text{if the observation is generated from the } k\text{-th component} \\ 0 & \text{else} \end{cases}$$

Carries from the k-th component

i.e.: $p(\underline{w} = (0, \ldots, \overset{k\text{-th}}{\downarrow} 1, \ldots 0)) = \pi_k$, or $p(\underline{w}) = \prod_{k=1}^{K} \pi_k^{w_k}$

$$\Rightarrow \underline{y} = \sum_{k=1}^{K} w_k \underline{y}_k, \quad \underline{y}_k \sim N(\underline{\mu}_k, \Sigma_k)$$

Conditional density of $\underline{Y} \mid \underline{w}$ is:

$$p\left(\underline{Y} \mid \underline{w}_k = 1\right) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}\left(\underline{Y}-\underline{\mu}_k\right)^T \Sigma_k^{-1} \left(\underline{Y}-\underline{\mu}_k\right)\right) \quad (*)$$

or

$$p(\underline{Y} \mid \underline{w}) = \prod_{k=1}^{K} [(*)]^{w_k}$$

and the JOINT DENSITY is:

$$p(\underline{Y}, \underline{w}) = \prod_{k=1}^{K} \left\{ \pi_k \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left\{-\frac{1}{2}\left(\underline{Y}-\underline{\mu}_k\right)^T \Sigma_k^{-1} \left(\underline{Y}-\underline{\mu}_k\right)\right\} \right\}^{w_k}$$

Why are these useful?

.) Flexible model to represent data that are multimodal

.) EASY TO SIMULATE FROM:

Pick one component (component) with prob π_k and then sample (sample) from $N(\mu_k, \Sigma_k)$

.) CLUSTERING (UNSUPERVISED LEARNING)

knowing that a data point is likely from one component is a way to cluster (i.e. group) the data.

$$p(w_k \mid Y) = \frac{p(Y, w_k)}{p(Y)} = \frac{p(w_k)\, p(Y \mid w_k)}{\sum_{l=1}^{K} p(w_l)\, p(Y \mid w_l)} =$$

$$= \frac{\pi_k\, p(Y \mid w_k = 1)}{\sum_{l=1}^{K} \pi_l\, p(Y \mid w_l = 1)}$$

$\Rightarrow$ LABEL $\hat{k} = \underset{k=1,\dots,K}{\arg\max}\; p(w_k \mid Y)$

To do this, I need to know the parameters $\underline{\vartheta}$; this is where we apply the parameter estimation techniques we have seen last time.

Let $\underline{Y}_1, \dots, \underline{Y}_N$ be i.i.d. with the same density as $\underline{Y}$. (GAUSSIAN MIXTURES with parameters $\underline{\vartheta}$). Then, we have LATENT VARIABLES (also called HIDDEN VARIABLES)

$\underline{W}_1, \dots, \underline{W}_N \sim \underline{W}$

They represent the component of the mixtures from which each data point comes from. Unfortunately, we do not ~~observe~~ OBSERVE them.

AIM: Find the MLE FOR $\underline{\vartheta}$

$$L(\vartheta; \underline{Y}, \underline{W}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left\{ \frac{\lambda_k}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (\underline{Y}_i - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{Y}_i - \underline{\mu}_k) \right] \right\}^{W_{ik}}$$

$$\Rightarrow \ell = \log(L) = \sum_{i=1}^{N} \sum_{k=1}^{K} W_{ik} \left[\log \lambda_k - \frac{p}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_k|) + \right.$$
$$\left. - \frac{1}{2} (\underline{Y}_i - \underline{\mu}_k) \Sigma_k^{-1} (\underline{Y}_i - \underline{\mu}_k) \right]$$

This would be easy to maximize if we knew $\underline{W}_1, \dots, \underline{W}_K$.

But we don't.

↳ By separating data coming from each component and estimating that parameters of each component separately

each →

$$\textcircled{+} \quad \hat{\lambda}_k = \frac{1}{N} \sum_{i=1}^{N} W_{ik}$$

estimating

⇒ Solve the problem using Expectation - Maximization algorithm

PARENTHESIS : INFORMATION THEORY AND KULLBACK-LEIBLER (KL) DIVERGENCE

INFORMATION THEORY STUDIES THE TRANSMISSION OF INFORMATION. the amount of information given by $\underline{Y}$ is measured by the entropy:

$$H(\underline{Y}) = -\int p(y) \log(p(y))\, dy$$

If we have a second distribution $q(\underline{Y})$, which is used to approximate $p(\underline{Y})$, we can define the relative entropy or KL DIVERGENCE between p and q: (i.e. AMOUNT OF INFORMATION LOST WHEN q IS USED TO APPROXIMATE p)

$$KL(p \| q) = -\int p(y) \log(q(y))\, dy + \int p(y) \log(p(y))\, dy =$$

$$KL = -\int p(y) \log\left[\frac{q(\underline{y})}{p(\underline{y})}\right] d\underline{y}$$

Note that: $KL(q \| p) \neq KL(p \| q)$

- By JENSON INEQUALITY

$$\int p(y) \log\left[\frac{q(\underline{y})}{p(\underline{y})}\right] \leq \log \underbrace{\int p(y) \frac{q(y)}{p(y)}\, dy}_{1} = 0$$

$$\Rightarrow \boxed{KL(p \| q) \geq 0}$$

CRUCIAL FACT WE ARE GOING TO USE.

- $KL(p \| q) = 0 \iff p(\underline{Y}) = q(\underline{Y})$ almost everywhere

EM ALGORITHM

Aim: $\max_{\underline{\vartheta}} \left[\log L(\underline{\vartheta}; \underline{Y}) \right] = \max_{\underline{\vartheta}} p(\underline{Y} | \underline{\vartheta})$

where $p(\underline{Y} | \underline{\vartheta}) = \sum_{\underline{W}} p(\underline{Y}, \underline{W} | \underline{\vartheta})$

↑ HARD TO MAXIMIZE

↑ EASY, BUT $\underline{W}$ IS "HIDDEN"

[redacted]

GENERAL IDEA OF THE ALGORITHM:

0.) START WITH SOME VALUE OF $\underline{\vartheta}$

1.) ESTIMATING THE PROBABILITY DISTRIBUTION OF $\underline{W}$, based on the current value of $\underline{\vartheta}$ (E-STEP)

2.) ESTIMATING A NEW VALUE OF $\underline{\vartheta}$, by MAXIMIZING

$$\sum_{W} p(\underline{Y} | \underline{\vartheta}, \underline{W}) p(\underline{W}) \qquad \text{(M-STEP)}$$

3.) ITERATE 1-2 UNTIL CONVERGENCE

More formally, we introduce a "trial" distribution $q(\underline{w})$ and we take advantage of the following identity:

$$\log p(Y|\theta) = \log\left(\frac{p(Y,\underline{w}|\theta)}{p(\underline{w}|Y,\theta)}\right) \overset{\text{(PRODUCT RULE)}}{=} \quad \forall\, \underline{w}$$

$$= \log\left(\frac{p(Y,\underline{w}|\theta)}{q(\underline{w})}\,\frac{q(\underline{w})}{p(\underline{w}|Y,\theta)}\right) =$$

$$= \log\left(\frac{p(Y,\underline{w}|\theta)}{q(\underline{w})}\right) + \log\left(\frac{q(\underline{w})}{p(\underline{w}|Y,\theta)}\right)$$

Expectation

IF we take the expectation w.r.t. $q(\underline{w})$:

$$\log p(Y|\theta) = \sum_{\underline{w}} q(\underline{w}) \log\left[\frac{p(Y,\underline{w}|\theta)}{q(\underline{w})}\right] + \;\cdot$$

$$\mathcal{Q}(q,\underline{\theta}) \qquad - \sum_{\underline{w}} q(\underline{w}) \log\left[\frac{p(\underline{w}|Y,\theta)}{q(\underline{w})}\right]$$

$$KL(q\,\|\,p) \geq 0$$

$$\Rightarrow \boxed{\log p(Y|\theta) > \mathcal{Q}(q,\theta)}$$

EM ITERATIONS:

- start $q^{(old)}, \underline{\theta}^{(old)}$
- $\max_{q} \mathcal{Q}(q, \underline{\theta}^{(old)}) = q^{new}$ (E-STEP)
- $\max_{\underline{\theta}} \mathcal{Q}(q^{new}, \underline{\theta}) = \underline{\theta}^{(new)}$ (M-STEP)

E-STEP: $Q(q, \theta^{(old)}) = \log p(Y|\theta^{(old)}) - KL(q \| p^{(old)})$

this is maximized when $KL(q \| p^{(old)}) = 0$

$$\Rightarrow \boxed{q^{new} = p(W|Y, \theta^{old})}$$

M-STEP: $\max_{\theta} Q(q^{new}, \theta)$

Note that $Q(q^{new}, \theta^{new}) \geq Q(q^{new}, \theta^{old})$,

then: $\log p(Y|\theta^{new}) = \underbrace{Q(q^{new}, \theta^{new})}_{\geq Q(q^{new}, \theta^{old})} + \underbrace{KL(q^{new} \| p^{new})}_{\geq 0}$

and $\log p(Y|\theta^{old}) = Q(q^{new}, \theta^{old}) + \underbrace{KL(q^{new} \| p^{old})}_{=0}$

$$\Rightarrow \boxed{\log p(Y|\theta^{new}) \geq \log p(Y|\theta^{old})}$$

i.e. each iteration of the E-M algorithm is guaranteed to increase the log-likelihood, i.e. it converges to a local maximum of $\log p(Y|\theta)$

M - STEP:

$$Q(q^{new}, \theta^{old}) = \sum_w q^{new}(w) \log \left[\frac{p(Y, w|\theta)}{q^{new}(w)} \right] =$$

$$= \underbrace{\sum_w q^{new}(w) \log p(Y, w|\theta)}_{\tilde{Q}(\underline{\theta}, \underline{\theta}^{OLD})} + \text{const}$$

EM FOR (GAUSSIAN) MIXTURES

- $$\log p(Y, w|\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} w_{ik} \left[\log \lambda_k - \frac{p}{2} \log \lambda - \frac{1}{2} \log |\Sigma_k| + \right.$$
$$\left. - \frac{1}{2} (\underline{Y}_i - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{Y}_i - \underline{\mu}_k) \right]$$

this following is from • pg2.

- $$p(\underline{w}|Y, \theta) = \prod_{i=1}^{N} p(\underline{w}_i | \underline{Y}_i, \theta)$$

AND $$p(w_{ik}=1 | \underline{Y}_i, \theta) = \frac{\lambda_k \, p(Y|w_k=1)}{\sum_{i=1}^{K} \lambda_i \, p(Y|w_i=1)} = \gamma_k(\underline{Y}_i)$$

$p(Y|w_k=1) \sim N(\mu_k, \Sigma_k)$

$\gamma_k(\underline{Y}_i)$ is called RESPONSABILITY of component k for the data point $\underline{Y}_i$.

responsibility

q^{new} from E-step

from M-Step last page (*) & subs. • & •• last page

$$\Rightarrow \quad \tilde{Q}\left(\underline{\theta}, \underline{\theta}^{old}\right) = \sum_{w} p\left(w \mid \underline{Y}, \underline{\theta}^{old}\right) \log p\left(\underline{Y}, w \mid \underline{\theta}^{new}\right) =$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_k\left(\underline{Y}_i\right) \left[\log \pi_k - \frac{p}{2} \log 2\pi - \frac{1}{2} \log \left|\Sigma_k\right| + \right.$$

$$\left. - \frac{1}{2} \left(\underline{Y}_i - \underline{\mu}_i\right)^T \Sigma_k^{-1} \left(\underline{Y}_i - \underline{\mu}_k\right) \right]$$

We can maximize separately in π_k and in μ_k, Σ_k:

$$\max_{\mu_k, \Sigma_k} \sum_{i=1}^{N} \gamma_k\left(\underline{Y}_i\right) \log \left\{ \frac{1}{(2\pi)^{p/2} \left|\Sigma_k\right|^{1/2}} \exp\left\{ -\frac{1}{2} \left(\underline{Y}_i - \underline{\mu}_k\right) \Sigma_k^{-1} \left(\underline{Y}_i - \underline{\mu}_k\right) \right\} \right)$$

Maximum is at:

$$\underline{\mu}_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_k\left(\underline{Y}_i\right) \underline{Y}_i, \qquad N_k = \sum_{i=1}^{N} \gamma_k\left(\underline{Y}_i\right)$$

Max. with Σ:

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_k\left(\underline{Y}_i\right) \left(\underline{Y}_i - \underline{\mu}_k\right) \left(\underline{Y}_i - \underline{\mu}_k\right)^T$$

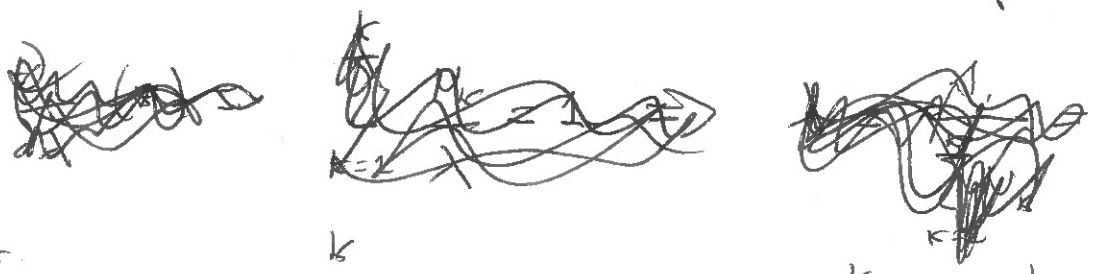
$$\max \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_k\left(\underline{Y}_i\right) \log \pi_k = \sum_{k=1}^{K} N_k \log \pi_k$$

with $\sum_{k=1}^{K} \pi_k = 1$. Using a Lagrange multiplier:

$$\max_{\pi} \left[\sum_{k=1}^{K} N_k \log\left(\pi_k\right) - \lambda \left(\sum_{k}^{K} \pi_k - 1 \right) \right] \Rightarrow$$

(*)

$$\Rightarrow \quad \frac{\partial}{\partial \pi_k}(*) = \frac{N_k}{\pi_k} - \lambda = 0 \Rightarrow \quad \pi_k = \frac{N_k}{\lambda}$$



$$\sum_{k=1}^{K} \pi_k = 1 \Rightarrow \sum_{k=1}^{K} \frac{N_k}{\lambda} = 1 \Rightarrow \lambda = \sum_{k=1}^{K} N_k = \sum_{k=1}^{K} \sum_{i=1}^{N} \gamma_k(x_i) = N$$

$$\Rightarrow \quad \boxed{\pi_k^{new} = \frac{N_k}{N}}$$

End of Week III

Week IV

From now it will be Prediction Problems.

4.) Linear Models

We start addressing the problem of prediction:

$$X_1, \dots, X_p \longrightarrow Y$$

EXPLANATORY VARIABLES — RESPONSE

We want to predict the response Y from the explanatory variables, i.e. $Y = f(X_1, \dots, X_p)$.

To LEARN the relationship f, we use a SAMPLE:

$$(X_{i1}, \dots, X_{ip}, Y_i), \quad i = 1, \dots, N$$

Gaussian Linear Model:

ϕ_i known, β_i unknown.

$$Y_i \mid \underline{X}_i = \underline{x}_i \overset{\text{i.i.d.}}{\sim} N(\mu_i, \sigma^2), \quad i = 1, \dots, N$$

where $\mu_i = \beta_1 \phi_1(\underline{x}_i) + \beta_2 \phi_2(\underline{x}_i) + \dots + \beta_r \phi_r(\underline{x}_i)$

MATRIX FORM:

$$\underline{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_r \end{bmatrix}$$

$$X = \begin{bmatrix} \phi_1(\underline{x}_1) & \cdots & \phi_r(\underline{x}_1) \\ \vdots & & \\ \phi_1(\underline{x}_N) & \cdots & \phi_r(\underline{x}_N) \end{bmatrix}$$

$$\Rightarrow \quad \underline{Y} = X\underline{\beta} + \underline{\varepsilon}; \quad \underline{\varepsilon} \sim N_N(\underline{0}, \sigma^2 I), \ \underline{\varepsilon} \text{ indep.}$$

$$\underline{Y} \sim N_N(X\underline{\beta}, \sigma^2 I)$$

$\underline{\varepsilon}$: errors

REMARK: More generally, we can use a linear model

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{\varepsilon}$ are independent but not necessarily Gaussian, most of the results will still hold.

LEAST SQUARES ESTIMATION

We can estimate $\underline{\beta}$ by minimizing the residual sum of squares:

$$S(\underline{\beta}) = \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} \phi_j(\underline{x}_i)\beta_j \right)^2 = (Y - X\beta)^T (Y - X\beta)$$

$$\frac{\partial S}{\partial \beta_k} = \frac{\partial}{\partial \beta_k} \sum_{i=1}^{N} \left[y_i - \sum_{j=1}^{p} \phi_j(\underline{x}_i)\beta_j \right]^2 =$$

$$= -2 \sum_{i=1}^{N} \left[y_i - \sum_{j=1}^{p} \phi_j(\underline{x}_i)\beta_j \right] \cdot \phi_k(\underline{x}_i)$$

$$\frac{\partial S}{\partial \underline{\beta}} = -2\, X^T (\underline{Y} - X\underline{\beta}) = 0$$

$$X^T (Y - X\underline{\beta}) = 0$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

IF $(X^T X)$ IS INVERTIBLE

IF $X^T X$ IS NOT INVERTIBLE → NOT UNIQUE SOLUTION.

We will see how to deal with this case with REGULARIZED METHOD

REMARK: IF $\underline{\varepsilon} \sim N_N(\underline{0}, \sigma^2 I)$, the MLE for $\underline{\beta}$ is the same as the LEAST SQUARES ESTIMATOR:

$$L(\underline{\beta}, \sigma^2; \underline{y}) = \frac{1}{(2\pi)^{N/2} |\sigma^2 I|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta)\right\}$$

$$\log L = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2 I| - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

$$\arg\max_{\beta} \log L = \arg\min_{\beta} (Y - X\beta)^T (Y - X\beta) =$$

$$= (X^T X)^{-1} X^T Y$$

- $\mathbb{E}[\hat{\beta}] = (X^T X)^{-1} X^T \mathbb{E}[Y] = (X^T X)^{-1} X^T X \beta = \beta$ UNBIASED ESTIMATOR
- $\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1} X^T \mathrm{Var}(Y) X (X^T X)^{-1} =$

$$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$$

$$\Rightarrow \hat{\beta} \sim N_p\left(\underline{\beta}, \sigma^2 (X^T X)^{-1}\right)$$

(SINCE LINEAR COMBINATION OF GAUSSIAN VARIABLES)

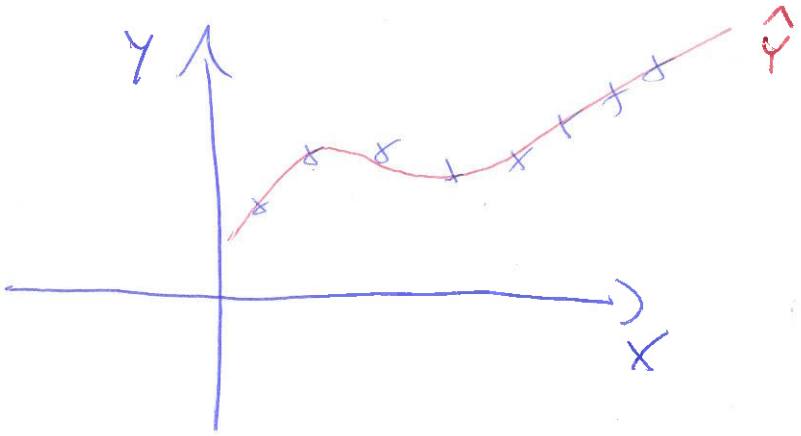
- THE MLE for σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{N}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

(LOOK FOR STRESS) but we know this is a BIASED estimator.

UNBIASED VERSION: $$s^2 = \frac{1}{N-p}(Y - X\hat{\beta})^T (Y - X\hat{\beta})$$

EXAMPLE : POLYNOMIAL REGRESSION:



$$Y_i = \mu_i + \varepsilon_i$$

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3$$

$$\phi_0(\underline{x}) = 1 \qquad \phi_1(\underline{x}) = x_i \qquad \phi_2(\underline{x}) = x_i^2 \qquad \phi_3(\underline{x}) = x_i^3$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 & x_N^3 \end{bmatrix}$$

PREDICTED VALUES: $\hat{Y} = \hat{\beta}_0 \phi_0(x) + \hat{\beta}_1 \phi_1(x) + \hat{\beta}_2 \phi_2(x) + \hat{\beta}_3 \phi_3(x)$

EXTENSIONS OF LINEAR MODELS (AND SPECIAL CASES)

- **MULTIVARIATE RESPONSE:** (when response Y is vector itself)

$$\underline{Y}_i \overset{i.i.d.}{\sim} N_m\left(\underline{x}_i^T B, \Sigma\right)$$

$$\underline{Y} = XB + E \longleftarrow \text{error}$$

$$E \sim \text{MATRIX VALUED NORMAL}\left(0_{nm}, \Sigma \otimes I_n\right)$$

LEAST SQUARES:

$$\hat{B} = \left(X^T X\right)^{-1} X^T Y$$

(same as many linear regressions)

Alternatively,

~~WEIGHTED~~ GENERALIZED LEAST SQUARES:

$$\hat{B} = \underset{B}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\underline{Y}_i - \underline{x}_i^T B\right)^T \Sigma^{-1} \left(\underline{Y}_i - \underline{x}_i B\right)$$

PROBLEM: Σ NEEDS TO BE ESTIMATED

POSSIBLE SOLUTIONS: ITERATIVE PROCEDURE

- $B^{(0)}, \Sigma^{(0)}$ — start values
- $\Sigma^{(j)} \to B^{(j+1)}$ — estimate B^{j+1} using Σ^j
- $B^{(j+1)} \to \Sigma^{(j+1)}$ — estimate Σ^{j+1} using B^{j+1}

- **CORRELATED RESPONSE:**

$$\underline{Y} \sim N_N\left(X\beta, \sigma^2 Z\right)$$

IF Z KNOWN, GENERALIZED LEAST SQUARES SOLUTION:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left(Y - X\beta\right)^T Z^{-1} \left(Y - X\beta\right) \Rightarrow \underline{\hat{\beta}} = \left(X^T Z^{-1} X\right)^{-1} X^T Z^{-1} \underline{Y}$$

- IF Σ UNKNOWN, AS ABOVE WE CAN USE AN ITERATIVE PROCEDURE.

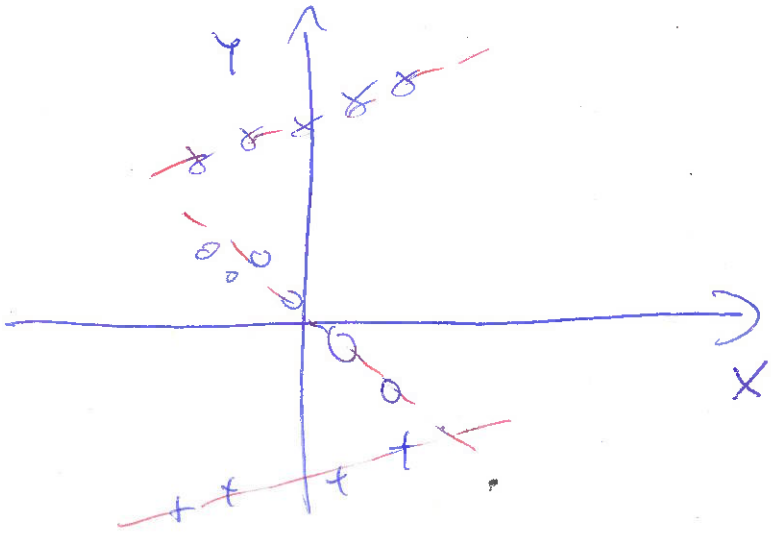
(POPULAR IN TIME SERIES AND SPATIAL STATISTICS)

but we need some assumptions on the structure of Σ.

- (We cannot estimate a N dimensional (dimensional) covariance with 1 observation)

EXAMPLE: TIME SERIES: $\text{cov}(\varepsilon_i, \varepsilon_s) = g(|i-s|)$

- INDICATOR VARIABLES VS RANDOM EFFECTS (Effects)



× GROUP 1
○ GROUP 2
+ GROUP 3

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 I_{1i} + \beta_3 I_{2i} + \beta_4 I_{1i} X + \beta_5 I_{2i} X + \varepsilon_i$$

INDICATOR VARIABLES

$$I_{1i} = \begin{cases} 1 & \text{IF } i \in \text{GROUP 2} \\ 0 & \text{ELSE} \end{cases} \qquad I_{2i} = \begin{cases} 1 & \text{IF } i \in \text{GROUP 3} \\ 0 & \text{ELSE} \end{cases}$$

⇒ THIS ALLOWS TO FIT INDIVIDUAL LINES FOR EACH GROUP.

HOWEVER, IF THE GROUPS ARE EXTRACTED FROM A LARGER POPULATION (EXAMPLES: SCHOOLS, HOSPITALS, COUNTIES, ETC..), IT IS BETTER TO USE RANDOM EFFECTS INSTEAD.

Linear Mixed Effects Models (Random effect)

$$\begin{cases} Y \mid u = \hat{u} \sim N(X\beta + Z\hat{u}, \sigma^2 I_n) \\ u \sim N(0, G) \end{cases}$$

— G is unknown
β is unknown

Random effect

$\hat{y}$

$$Y = X\beta + Zu + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I_n), \quad u \sim N(0, G)$$

$$\Rightarrow Y \sim N(X\beta, \Sigma), \quad \Sigma = ZGZ^T + \sigma^2 I_n$$

$$\Rightarrow \hat{\beta} = (X\Sigma^{-1}X)^{-1} X^T \Sigma^{-1} Y \leftarrow (\text{Generalised Least Sq using})$$

$\hat{\Sigma}$ = numerically / numerically

(REML ALGORITHM)

↓

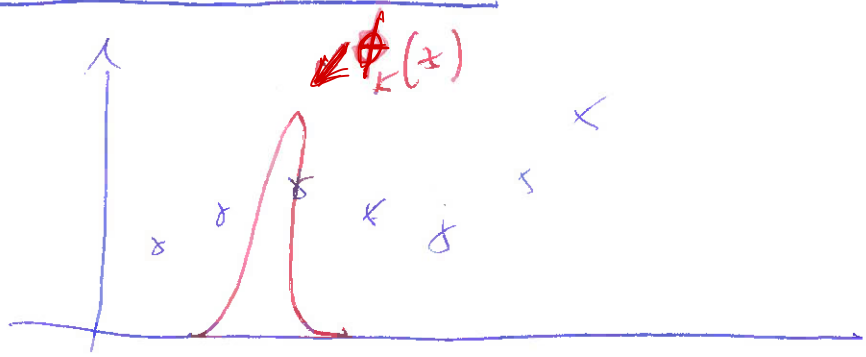
Restricted Max. likelihood Algorithm (REML)

End of Week IV

Week V

(last week material continues)

- **Local Basis Functions**



$$Y_i = \sum_{j=1}^{p} \beta_j \phi_j(x) + \varepsilon_i$$

EXAMPLES: SPLINES, WAVELETS, LOCAL POLYNOMIAL

- **Additive Models**

$$Y_i = \sum_{j=1}^{p} f_j(X_j) + \varepsilon_i \quad , i = 1, \ldots, n$$

↳ TO BE ESTIMATED SOME HOW, FOR EXAMPLE:

$$f_j(x_j) = \sum_{k=1}^{R} \beta_{kj} \phi_k(x_j)$$

$$\Rightarrow Y_i = \sum_{j=1}^{p} \sum_{k=1}^{R} \beta_{kj} \phi_k(x_j) + \varepsilon_i$$

IN PRACTICE, FITTED (fonctions) VIA BACKFITTING ITERATIVE (itératif) PROCEDURE; OR PENALIZED METHODS (Régularisés)

5) Model Selection and Regularization

(New material for this week)

How do we choose the model to fit?

How to choose which and how many functions $\phi_s(\underline{x})$ to use?

REMARK: . Whenever possible, scientific knowledge about the variables of interest needs to be used to restrict as much as possible the set of models to be considered.

However, we will focus on how to use the data to select among a set of model.

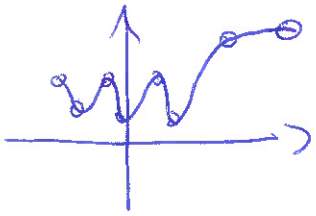
EXAMPLE: Polynomial Regression: $\underline{Y} \sim N_N(\underline{\mu}, \sigma^2 I_N)$

$$\mu_i = \beta_0 + \beta_1 x_i + \dots + \beta_p x_i^p$$

How should we choose p (degree of the polynomial)?

Note: $p = N-1 \Rightarrow S(\underline{\beta}) = 0$ (INTERPOLATING POLYNOMIAL)

However, the model is useless because it is adapted to the error in our data $\Rightarrow$ EXAMPLE of OVERFITTING:



TWO ASPECTS OF MODEL SELECTION:

1) • CRITERION TO COMPARE TWO MODELS

2) • SEARCH OF THE MODEL SPACE

For what concerns 2), if the model space is small you can compute your criterion for all models and pick the best one. If the model space is large, this can be NOT computationally feasible. A popular alternative is to use STEPWISE algorithm, which add or remove one predictor at the time from the model and they stop when the chosen criterion cannot be improved by adding or removing any term. It is a GREEDY approach, no guarantees to find the best model overall.

About 1), we are going to see a gallery of used criteria:

(i) – SPLIT DATA IN TRAINING / TEST SET

You FIT your models using the training set and the criterion is the prediction error on the data in the test set.

For example, $i=1,\dots,n$ → TRAINING SET

$i=n+1,\dots,N$ → TEST SET

Use $Y_1,\dots,Y_n$ → estimate $\hat{\beta}$

Choose the model which minimize $\sum_{i=n+1}^{N} (Y_i - \underline{x}_i^T \hat{\beta})^2$

DRAWBACK: • IT WASTES SOME DATA, LESS ACCURATE ESTIMATES

• THE ESTIMATE OF THE PREDICTION ERROR IS ALSO HIGHLY VARIABLE (SMALL TEST SAMPLE)

(ii) — CROSS-VALIDATION:

You repeat the partition between training and testing set multiple times and look at the average prediction error

Special cases: → as opposed to once in part (i) & then take average

(a) K-FOLD CROSS VALIDATION:

- Split the data in K set
- for k=1:K:
  - fit the model without the data in the k-th set
  - Compute $e_k = \frac{1}{N_k} \sum_{i \in N_k} (Y_i - \hat{Y}_i^{(-k)})^2$

    where N_k is the set of indices in the k-th partition and $\hat{Y}_i^{(-k)}$ is the prediction of the model fitted without the k-th set of data.
- Choose the model that minimizes $\frac{1}{K} \sum_{k=1}^{K} e_k$ ← average of prediction errors

(b) · LEAVE-1-OUT CV:

- for i=1:N
  - fit the model without the i-th observation
  - compute $e_i = (Y_i - \hat{Y}_i^{(-i)})^2$
- Choose the model that minimizes $\frac{1}{N} \sum_{i=1}^{N} e_i^2$

ADVANTAGES: You extract information from all the data

DRAW BACKS: - COMPUTATIONALLY EXPENSIVE

- YOU CAN STILL OVERFIT YOUR MODEL TO YOUR SPECIFIC DATASET.

AKAIKE INFORMATION CRITERION (AIC)

This is one example of criteria that try to balance the prediction error and the complexity of the model:

$$AIC = 2p - 2 \log L\left(\hat{\beta}, \hat{\sigma}^2; \underline{Y}\right)$$

↳ p is the number of parameters in the model

OTHER INFORMATION CRITERIA use different penalties:

BIC: $\log(N)\, p$

DIC: ---

- For large N, AIC:
  - approximate the LEAVE-1-OUT CV error
  - the difference in AIC between two models (difference in) is an estimate of the KL divergence between these models and the true model.

- Choose model that minimizes AIC.

REGULARIZATION

Instead of selecting a model, we can use the largest possible model but SHRINKING the estimated coefficients to control their variance (and thus avoid overfitting).

• RIDGE REGRESSION:

$$\hat{\beta}_R = \underset{\beta}{\operatorname{argmin}} \; (\underline{Y} - X\underline{\beta})^T (\underline{Y} - X\underline{\beta}) =$$

$$\underline{\beta}^T\underline{\beta} \le c$$

→ SHRINKING CONSTRAINT

LAGRANGIAN MULTIPLIER $= \underset{\beta}{\operatorname{argmin}} \left\{ (\underline{Y} - X\underline{\beta})^T (\underline{Y} - X\underline{\beta}) + \lambda \underline{\beta}^T \underline{\beta} \right\}$

$\tilde{S}(\underline{\beta};\lambda)$

Now minimize $\tilde{S}(\underline{\beta};\lambda)$,

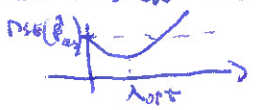
$$\frac{\partial \tilde{S}}{\partial \underline{\beta}} = -2X^T\underline{Y} + 2X^TX\underline{\beta} + 2\lambda\underline{\beta} = 0$$

$$\Rightarrow \quad (X^TX + \lambda I)\underline{\beta} = X^T\underline{Y}$$

$$\boxed{\hat{\beta}_R = (X^TX + \lambda I)^{-1} X^T\underline{Y}}$$

REMARKS:

- Above, $c = \hat{\beta}_R^T \hat{\beta}_R$; λ is a TUNING PARAMETER TO BE CHOSEN.
- $\hat{\beta}_R$ is well defined even if X^TX is NOT INVERTIBLE.
- $\lambda = 0 \Rightarrow \hat{\beta}_R = \hat{\beta}_{OLS}$; $\lambda \to +\infty \Rightarrow \hat{\beta}_R \to \underline{0}$
- It can be shown that the MEAN SQUARED ERROR AS FUNCTION OF λ follows

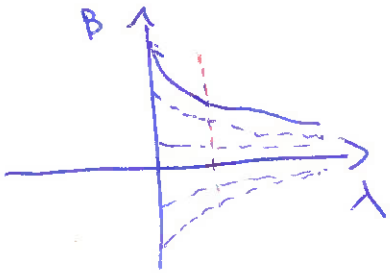


Remarks: - The shrinking happens equally on all coefficients, so we need to standardized the variables ~~first~~ first.

- How to choose λ:

(i) Cross-validation

(ii) Ridge trace



minimum λ that "stabilizes" the estimates

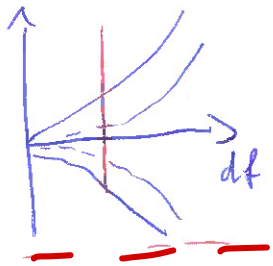
(iii) Approximation of λ_{opt} given some estimate of σ^2

Once λ is chosen then check:

- Effective degrees of freedom: (Effective degrees of freedom of model.)

$$\hookrightarrow df(\lambda) = \text{trace}\left(X\left(X^TX+\lambda I\right)^{-1}X^T\right)$$

since $\hat{Y} = X\left(X^TX+\lambda I\right)^{-1}X^TY$, it is a measure of the dimension of the space to which fitted values belong.



If $\lambda \to \infty$, $df(\lambda) \to 0$

$\lambda \to 0$, $df(\lambda) \to$ original no. of parameters.

---

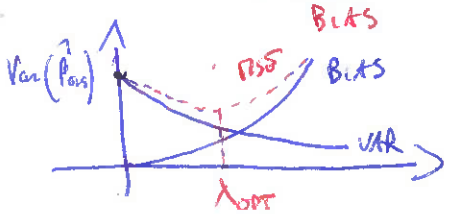
* Aside: Bias-variance trade-off

$W = \left(X^TX+\lambda I\right)^{-1}$

$$MSE\left(\hat{\beta}_R\right) = \left\| E\left[\hat{\beta}_R - \beta\right]\right\|_2^2 + \text{trace}\left(Var\left(\hat{\beta}_R\right)\right) =$$

$$= \lambda^2 \beta^T W^2 \beta + \sigma^2 \text{trace}\left(W(X^TX)W\right)$$

Bias (squared) — Variance



$\lambda = 0$: Bias $= 0$, $Var = Var\left(\hat{\beta}_{OLS}\right)$ (a2 + ∞)

$\lambda \to +\infty$: Bias $\to +\infty$, $Var \to 0$

(End of Week II)

Week VI

(last week material continued)

LASSO:

There is no particular reason to shrink the coefficients based on the 2-norm. The LASSO estimator is instead.

$$\hat{\beta}_L = \arg\min_{\beta} (\underline{Y} - X\beta)^T (\underline{Y} - X\beta) =$$

$$\sum_j |\beta_j| \leq \tau \quad \leftarrow \text{constraint}$$

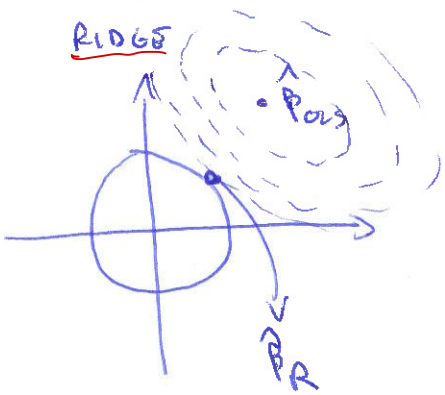
$$\|\beta\|_1$$

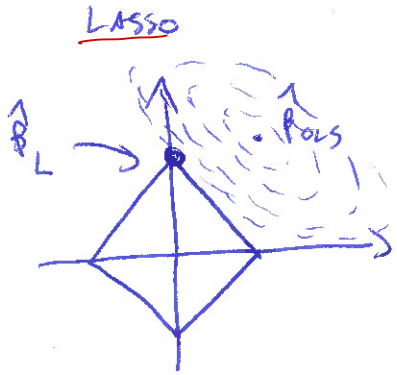
$$= \arg\min_{\beta} \left\{ (Y - X\beta)^T (Y - X\beta) + \lambda \sum_j |\beta_j| \right\}$$

No explicit solution, it needs to be solved numerically (LARS algorithm) (Also called Norm 1 (Lasso))

REMARKS:

- Again, variables need to be standardised first.
- $\lambda \to 0$, $\hat{\beta}_L \to \hat{\beta}_{ols}$; $\lambda \to +\infty$, $\hat{\beta}_L \to 0$
- VISUAL COMPARISON IN 2D:





Minimum occurs at boundary of ellipse & circle or ellipse & line.

LASSO ESTIMATOR favors (favours) SPARSITY.

- Choice of λ based on similar strategies as in the RIDGE regression.
- In statistical software, often reparameterized as $s = \frac{\tau}{\|\hat{\beta}_{ols}\|_1}$, so that $0 < s < 1$.

- It is possible to play around further with the penalty, tailoring it to the problem and basis function used.

EXAMPLES: . Smoothing splines:

$$\hat{f} = \sum_{i=1} \hat{\beta}_i f_i(x)$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{n} \left(y_i - \hat{f}(x_i)\right)^2 + \lambda \int_a^b \left[\hat{f}''(t)\right]^2 dt$$

- LASSO is equivalent to soft-thresholding

$$\hat{\beta}_{L,j} = \hat{\beta}_{OLS,j} \max\left(0, 1 - \frac{\lambda}{|\hat{\beta}_{OLS,j}|}\right)$$

IF THE PREDICTORS ARE ORTHOGONAL
(EXAMPLE: WAVELETS OR FOURIER BASIS)

- BEST SUBSET REGRESSION:

$$\hat{\beta}_0 = \underset{\beta}{\operatorname{argmin}} (Y - X\beta)^T (Y - X\beta) + \lambda \|\beta\|_0$$

where $\|\beta\|_0 =$ NUMBER OF COEFFICIENTS $\neq 0$

- Lasso does shrinkage & selection.

6) BAYESIAN LINEAR REGRESSION

(New material)

← regularisation method

We will now consider a Bayesian approach to estimate the parameters of linear models. As mentioned before, a Bayesian approach needs us to specify a PRIOR for the parameters in the model.

MODEL:

$$\begin{cases} Y \mid X, \beta \sim N_N(X\beta, \sigma^2 I) \\ \theta = (\beta, \sigma^2) \sim \text{PRIOR DISTRIBUTION } p(\theta) \end{cases}$$

In general, you will use Bayes theorem + numerical method to get a (sample from) the posterior distribution, from which you can get estimates for β and σ^2 by minimizing the chosen LOSS FUNCTION (as seen in one of the numerical examples in the tutorial).

As an example / exercise, we consider here the simple case where:

- σ^2 IS KNOWN
- $\beta \sim N_p(\underline{0}, \sigma_\beta^2 I_p)$, i.e. zero mean and ISOTROPIC (same variance for all coefficients and no correlations, hence independent)

} Prior

[REMARK: the zero-mean prior is not restrictive, you can always transform the variable appropriately. For example, if $\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \sim N_2\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \sigma_\beta^2 I_2\right)$, you can transform to:

$$Y_i^* = Y_i - \alpha - \beta X_i = \underbrace{(\beta_0 - \alpha)}_{\beta_0^*} + \underbrace{(\beta_1 - \beta)}_{\beta_1^*} X_i + \varepsilon_i$$

$$\Rightarrow \begin{pmatrix} \beta_0^* \\ \beta_1^* \end{pmatrix} \sim N_2\left(\underline{0}, \sigma_\beta^2 I_2\right)$$

]

We need to find posterior distribution. To do this we find joint density of Y & β & by conditioning find posterior dist. i.e $\theta|y$.

JOINT DENSITY OF $\underline{Y}$ AND $\underline{\beta}$:

$$p\left(\underline{Y},\underline{\beta}\mid X,\sigma^2,\sigma_\beta^2\right) = p\left(\underline{Y}\mid X,\underline{\beta},\sigma^2\right)p\left(\underline{\beta}\mid\sigma_\beta^2\right) =$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}}\frac{1}{(2\pi\sigma_\beta^2)^{\frac{p}{2}}}\exp\left[-\frac{1}{2\sigma^2}(\underline{Y}-X\underline{\beta})^T(\underline{Y}-X\underline{\beta})-\frac{1}{2\sigma_\beta^2}\underline{\beta}^T\underline{\beta}\right] =$$

$$\propto \exp\left(-\frac{1}{2}\begin{bmatrix}\underline{Y}\\ \underline{\beta}\end{bmatrix}^T\begin{bmatrix}\frac{1}{\sigma^2}I_N & -\frac{1}{\sigma^2}X\\ -\frac{1}{\sigma^2}X^T & \frac{1}{\sigma_\beta^2}I_p+\frac{1}{\sigma^2}X^TX\end{bmatrix}\begin{bmatrix}\underline{Y}\\ \underline{\beta}\end{bmatrix}\right)$$

$$\Rightarrow \begin{pmatrix}\underline{Y}\\ \underline{\beta}\end{pmatrix} \sim N_{N+p}\left(\underline{0}, \underbrace{\Lambda^{-1}}_{\Sigma}\right)$$

FROM THE PROPERTIES WE HAVE SEEN FOR THE MULTIVARIATE NORMAL:

$$\boxed{\underline{\beta}\mid\underline{Y}\sim N_p\left(\tilde{\underline{\beta}},\tilde{\Sigma}\right)}, \text{ where}$$

$$\tilde{\underline{\beta}} = \Sigma_{21}\Sigma_{11}^{-1}\underline{Y} = -\left(\Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)\Lambda_{12}\underline{Y} =$$

$$= -\Lambda_{22}^{-1}\Lambda_{21}\underline{Y} = \left(\frac{1}{\sigma_\beta^2}I_p+\frac{1}{\sigma^2}X^TX\right)^{-1}\frac{1}{\sigma^2}X^T\underline{Y} =$$

$$= \boxed{\left(X^TX+\frac{\sigma^2}{\sigma_\beta^2}I_p\right)^{-1}X^T\underline{Y}}$$

AND $\tilde{\Sigma} = \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \Lambda_{22}^{-1} = \boxed{\left(\frac{1}{\sigma_\beta^2}I+\frac{1}{\sigma^2}X^TX\right)^{-1}}$

posterior dist

- **MAP ESTIMATOR:**

$$\hat{\beta} = \mathbb{E}[\beta \mid Y] = \left(X^T X + \frac{\sigma^2}{\sigma_\beta^2} I_p \right)^{-1} X^T Y$$

The more error in data the more regularisation I need to add & more uncertainty of prior the estimator will be close to Ridge estimator. More certain of prior & less of data, the more we are shrinking the parameters.

λ : EQUIVALENT TO THE RIDGE REGRESSION ESTIMATOR

- **PREDICTIVE DISTRIBUTION:**

We now want to predict the response Y^* at a new set of values of the predictors $\underline{x}^* = (\phi_1(\underline{x}), \dots, \phi_p(\underline{x}))$. In a Bayesian setting, we can compute the PREDICTIVE DISTRIBUTION:

$$\boxed{p(Y^* \mid \underline{x}^*, \underline{y}, X, \sigma^2, \delta_\beta^2) = \int p(y^* \mid \underline{x}^*, \underline{\beta}, \sigma^2) p(\underline{\beta} \mid Y, X, \sigma^2, \delta_\beta^2) d\beta}$$

This is an example of GAUSSIAN LINEAR MODEL:

Assume $\underline{z} \sim N_N(\underline{\mu}, \Lambda^{-1})$ AND $\underline{U} \mid \underline{z} = N_N(A\underline{z} + \underline{b}, L^{-1})$, then the marginal density of $\underline{U}$ is also normal and

$$\mathbb{E}[\underline{U}] = \int \mathbb{E}[\underline{U} \mid \underline{z} = z] p(z) dz = \int (Az + b) p(z) dz = A\underline{\mu} + \underline{b}$$

$$Cov(\underline{U}) = \mathbb{E}[(\underline{U} - A\mu - b)(\underline{U} - A\mu - b)^T] =$$

SUB → $\underline{U} = A\underline{z} + \underline{b} + \underline{\varepsilon}$, $\underline{\varepsilon} \sim N(0, L^{-1}) \perp \underline{z}$ (independent)

$$= \mathbb{E}[(Az + b + \varepsilon - A\mu - b)(Az + b + \varepsilon - A\mu - b)^T] =$$

$$= \mathbb{E}[A(z - \mu)(z - \mu)^T A^T] + \mathbb{E}[\varepsilon \varepsilon^T] =$$

$$Cov(\underline{U}) = A \Lambda^{-1} A^T + L^{-1}$$

BACK TO THE **PREDICTIVE DISTRIBUTION**:

$$\begin{cases} Y^* \mid \underline{x}^*, \underline{y}, X, \sigma^2, \sigma_\beta^2 \sim N\left(\underline{x}^{*T}\hat{\beta},\ \sigma^2 I + \underline{x}^{*T}\tilde{\Sigma}\,\underline{x}^*\right) \\ \text{SINCE: } \beta \mid y, X, \sigma^2, \sigma_\beta^2 \sim N_p\left(\hat{\beta}, \tilde{\Sigma}\right) \\ \text{AND } Y^* \mid x^*, \beta, \sigma^2 \sim N\left(\underline{x}^{*T}\beta, \sigma^2\right) \end{cases}$$

($\hat{\beta}$: MAP ESTIMATOR; $\underline{x}^{*T}\tilde{\Sigma}\,\underline{x}^*$: UNCERTAINTY IN β)

- **BAYESIAN MODEL SELECTION**:

We now have candidate models $M_1, \dots, M_L$ ($\Pi_1, \dots, \Pi_L$), each with a prior probability $p(\Pi_i)$ of being the "true" model.

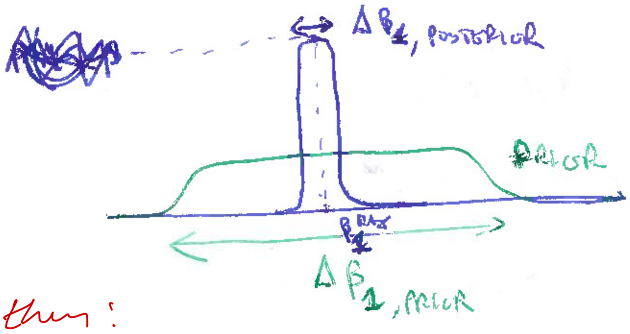
A natural way to do model selection is to look at posterior probability

$$p(\Pi_i \mid \underline{y}) \propto p(\Pi_i)\, p(y \mid \Pi_i)$$

($p(y \mid \Pi_i)$: MODEL EVIDENCE)

$$p(Y \mid \Pi_i) = \int p(\underline{y} \mid \Pi_i, \beta_i)\, p(\beta \mid \Pi_i)\, d\beta$$

Let's try to understand the model evidence in a simple scenario:



In this case then:

$$\text{Then, } p(Y \mid \Pi_i) \simeq p(\underline{Y} \mid \Pi_i, \beta_1^{MAX})\, \frac{\Delta\beta_{1,\text{POSTERIOR}}}{\Delta\beta_{1,\text{PRIOR}}}$$

($p(Y \mid \Pi_i)$: Model evidence; $p(\underline{Y} \mid \Pi_i, \beta_1^{MAX})$: Likelihood of observation max posteriore.)

Taking the log:

$$\log p(y \mid \Pi_i) \approx \log p(\underline{y} \mid \Pi_i, \hat{\beta}_{\Pi_i}^{MAP}) + \log\left(\frac{\Delta\beta_{1,\text{posterior}}}{\Delta\beta_{1,\text{prior}}}\right)$$

$\log p(\underline{y} \mid \Pi_i, \hat{\beta}_{\Pi_i}^{MAP})$: log-likelihood in a frequentist setting

$\log\left(\frac{\Delta\beta_{1,\text{posterior}}}{\Delta\beta_{1,\text{prior}}}\right)$: penalization term

If we consider many parameters $\beta_1, \ldots, \beta_p$ with the same ratio $\frac{\Delta \text{posterior}}{\Delta \text{prior}}$,

$$\log p(y \mid \Pi_i) \approx \log p(y \mid \Pi_i, \hat{\beta}^{MAP}) + p \log\left(\frac{\Delta\beta_{1,\text{posterior}}}{\Delta\beta_{1,\text{prior}}}\right)$$

$\log p(y \mid \Pi_i)$: model evidence

$\log p(y \mid \Pi_i, \hat{\beta}^{MAP})$: model fit

$p \log\left(\frac{\Delta\beta_{1,\text{posterior}}}{\Delta\beta_{1,\text{prior}}}\right)$: penalization for complexity

However, a better approach would be to do model averaging instead of model selection:

$$p(y^* \mid \underline{x}^*, \underline{y}, X, \sigma^2, \delta_\beta^2) = \sum_{i=1}^{L} p(y^* \mid \underline{x}^*, \underline{y}, X, \sigma^2, \delta_\beta^2, \Pi_i)\, p(\Pi_i \mid y)$$

Remark:

- In theory, if the true model is one of $\Pi_1, \ldots, \Pi_L$, Bayesian approach avoids overfitting even when using all the data to fit the model.
- In practice, there are a couple of issues:
  - prior specification
  - true model not in the set $\Pi_1, \ldots, \Pi_L$
- It is better to use a train/test set approach to check the model: posterior predictive checks

End of Week VI

Week VII

- **Empirical Bayes** → Get prior from data

Until now, we have assumed σ^2 known, which is unrealistic.

We can of course put a prior on σ^2 as well.

For example:

$$\frac{1}{\sigma^2} \sim \text{Gamma}(\alpha, \beta)$$

$$\beta \sim N_p(0, \sigma_\beta^2 I) \quad \text{independently}$$

(We will see this example in the tutorial)

More complicated priors can be used for (β, σ^2), and numerical methods (MCMC) can be used to sample from the posterior.

However, you still need to choose the hyperparameters $(\alpha, \beta, \sigma_{\beta}^2)$

An alternative is to use an approximation, known as

- **Empirical Bayes or Evidence Approximation.**

Generally speaking, empirical Bayes refers to the fact that the priors (or some of its parameters) are estimated from the data.

Note that this violates the standard Bayes procedure, where the prior needs to be known in advance.

We will again consider a special scenario:

$$\begin{cases} Y \mid \beta, \sigma^2 \sim N_N(X\beta, \sigma^2 I_N) \\ \beta \sim N_p(0, \sigma_\beta^2 I_p) \\ \sigma^2 \sim p(\sigma^2) \rightarrow \text{PRIOR on } \sigma^2 \\ \sigma_\beta^2 \sim p(\sigma_\beta^2) \rightarrow \text{HYPER PRIOR on } \sigma_\beta^2 \end{cases}$$

Now, the predictive distribution is:

$$p(y^*|y) = \iiint p(y^*|\beta,\sigma^2)\, p(\beta|y,\sigma_\beta^2,\sigma^2)\, p(\sigma^2,\sigma_\beta^2|y)\, d\beta\, d\sigma^2\, d\sigma_\beta^2$$

since we need to integrate w.r.t. σ^2 and σ_β^2.

APPROXIMATION (EMPIRICAL BAYES) 1: The posterior $p(\sigma^2,\sigma_\beta^2|y)$ is SHARPLY PEAKED around $\hat{\sigma}^2, \hat{\sigma}_\beta^2$. ↳ max value

↳ (assumption ①)

$$\Rightarrow p(y^*|y) \approx \int p(y^*|\beta,\hat{\sigma}^2)\, p(\beta|y,\hat{\sigma}^2,\hat{\sigma}_\beta^2)\, d\beta$$

(Same expression of the predictive distribution as with σ^2 and σ_β^2 known, but plugging in $\hat{\sigma}^2$ and $\hat{\sigma}_\beta^2$ instead)

So what remains is only to find $\hat{\sigma}^2, \hat{\sigma}_\beta^2$.

APPROXIMATION (EMPIRICAL BAYES) 2: $p(\sigma^2,\sigma_\beta^2) \propto$ constant (IMPROPER PRIOR) ← assumption ②

(Can be obtained by maximising marginal likelihood as below)

$$\Rightarrow (\hat{\sigma}^2,\hat{\sigma}_\beta^2) = \underset{\sigma^2,\sigma_\beta^2}{\operatorname{argmax}}\; p(\sigma^2,\sigma_\beta^2|y) =$$

$$= \underset{\sigma^2,\sigma_\beta^2}{\operatorname{argmax}}\; p(y|\sigma^2,\sigma_\beta^2)\, p(\sigma^2,\sigma_\beta^2) =$$

$p(\sigma^2,\sigma_\beta^2)$ constant

$$= \underset{\sigma^2,\sigma_\beta^2}{\operatorname{argmax}}\; p(y|\sigma^2,\sigma_\beta^2) = \underset{\sigma^2,\sigma_\beta^2}{\operatorname{argmax}} \int p(y|\beta,\sigma^2)\, p(\beta|\sigma_\beta^2)\, d\beta$$

$d\beta$

$p(y|\sigma^2,\sigma_\beta^2)$: MARGINAL LIKELIHOOD → likelihood of my data given only some parameters as β is not here.

Marginal likelihood

The marginal likelihood $p(y \mid \sigma^2, \delta_p^2)$ can be maximised numerically, either directly, or using an EM algorithm with β as latent variable (see problem sheet 7).

- What is interesting to note is this allows us to estimate $\frac{\hat{\sigma}^2}{\hat{\delta}_p^2}$ from the data, without doing any cross-validation.
  - → REGULARISATION PARAMETER penalty IN THE BAYESIAN MAP ESTIMATOR
- This procedure is also called Evidence approx. in Machine learning.

Gaussian Process Regression

This is an extension of Bayesian linear model, where we do not specify the expression of the conditional mean $M(\underline{x})$.

Indeed, in a Bayesian linear model we have: $Y_i(\underline{x}_i) = M(\underline{x}_i) + \varepsilon_i$, with $M(\underline{x}) = \underline{\beta}^T \underline{\phi}(\underline{x})$ and prior $\underline{\beta} \sim N_p(\underline{0}, \sigma_\beta^2 I)$.

For N observations,

$$\underline{M} = \begin{bmatrix} M(\underline{x}_1) \\ \vdots \\ M(\underline{x}_N) \end{bmatrix} \sim N_N(0, K) \quad \text{a priori}.$$

because $\underline{M} = X\underline{\beta}$, so it is a Gaussian (linear function of a Gaussian r.v.) with $\mathbb{E}[\underline{M}] = X\,\mathbb{E}[\underline{\beta}] = 0$ and $K = \mathrm{Var}(M) = X\,\mathrm{Var}(\beta)\,X^T = \sigma_\beta^2 X X^T$.

Note that $K_{ij} = \sigma_\beta^2 \underline{\phi}(\underline{x}_i)\underline{\phi}(\underline{x}_j)^T$, which is a function of the evaluations of the predictors in observations i and j.

Therefore, we can redefine the model by using the prior

$$\underline{M} \sim N_N(\underline{0}, K)$$

this motivates us to generalize this model by changing expectation and variance, i.e. defining a prior directly on the space of the functions $\underline{M}(x)$.

A GAUSSIAN PROCESS is a probability distribution over functions $H(x)$, such that the evaluation of H at an arbitrary set of points $x_1, \dots, x_N$ has distribution

$$\begin{bmatrix} H(x_1) \\ \vdots \\ H(x_N) \end{bmatrix} \sim \mathcal{N}_N \left(m(x), K \right),$$

(GRAM MATRIX)

with $K_{ij} = k(\underline{x}_i, \underline{x}_j)$

↳ KERNEL FUNCTION

the kernel function $k(\cdot,\cdot)$ needs to be such that K (GRAM MATRIX) is positive semi-definite $\forall\ \underline{x}_1, \dots \underline{x}_N$. & Symmetric.

OBTAINING VALID KERNELS:

Useful rules to construct valid kernels:

- SUM: If $k_1(\underline{x}, \underline{x}')$ and $k_2(\underline{x}, \underline{x}')$ are valid kernels, then $k(\underline{x}, \underline{x}') = k_1(\underline{x}, \underline{x}') + k_2(\underline{x}, \underline{x}')$ is a valid kernel.

PROOF: $K_{ij} = k(\underline{x}_i, \underline{x}_j) = k_1(\underline{x}_i, \underline{x}_j) + k_2(\underline{x}_i, \underline{x}_j)$

$\Rightarrow K = K_1 + K_2$

$v^T K v = v^T K_1 v + v^T K_2 v \geq 0$

($v^T K_1 v \geq 0$, $v^T K_2 v \geq 0$)

- **PRODUCT**: if $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$ are valid kernels, then $k(\underline{x},\underline{x}^*) = k_1(\underline{x},\underline{x}^*)\,k_2(\underline{x},\underline{x}^*)$ is valid.

PROOF: Let $H_1(\underline{x})$ be a sample from GP (Gaussian process) $(H_1 \mid \underline{0}, k_1)$

$H_2(\underline{x})$ " " " $GP(H_2 \mid 0, k_2)$, independently

$$k_{ij} = \mathbb{E}\left[H_1(\underline{x}_i) H_2(\underline{x}_i) H_1(\underline{x}_j) H_2(\underline{x}_j)\right] =$$

$H_1 \perp H_2$

$$= \mathbb{E}\left[H_1(\underline{x}_i) H_1(\underline{x}_j)\right] \mathbb{E}\left[H_2(\underline{x}_i) H_2(\underline{x}_j)\right] = (k_1)_{ij}(k_2)_{ij}$$

Thus is a covariance matrix, so $k(\cdot,\cdot)$ is a valid kernel (because K is positive semi-definite)

- ~~Ex~~

- **Variable transformation**: $k_1(\cdot,\cdot)$ valid,

$$k(x, x^*) = k_1(f(x), f(x^*)) \text{ is}$$

VALID using some transformation f in f.

- $k(x,x^*) = g(x)\,g(x^*)$ is VALID

- $k(x,x^*) = \exp\left(k_1(x,x^*)\right)$ is VALID

- $k(x,x^*) = c\left(k_1(x,x^*)\right)^{\alpha}$, $\alpha > 0$, $c > 0$, is VALID

- **EXAMPLES**:

  - GAUSSIAN KERNEL: $k(\underline{x},\underline{x}^*) = \exp\left(-\dfrac{\|x - x^*\|^2}{2\sigma^2}\right)$

  - POLYNOMIAL KERNEL: $k(\underline{x},\underline{x}^*) = \left(\underline{x}^T \underline{x}^*\right)^{\alpha}$

• GP REGRESSION

Now assume $Y_i = f_i + \varepsilon_i$, $f_i = f(\underline{x}_i)$, with

$$f(\underline{x}) \sim GP(\underline{\mu} \mid \underline{0}, K)$$
$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

mean = 0; kernel fn or covariance K

$$\Rightarrow \underline{Y}_n \mid \underline{f} \sim N_n(\underline{f}, \sigma^2 I)$$
$$\underline{f} \sim N_N(0, K) \leftarrow \text{Prior}$$

Thus is again an example of GAUSSIAN LINEAR MODEL (see last lecture!) then the marginal distribution of $\underline{Y}$ is GAUSSIAN

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid f]] = 0$$

$$Var(Y) = I K I^T + \sigma^2 I = K + \sigma^2 I$$

$$\Rightarrow \underline{Y} \sim N_N(\underline{0}, K + \sigma^2 I)$$

PREDICTIVE DISTRIBUTION:

$$Y_0 \mid \underline{Y} \sim N\left(\underline{k}_0^T (K + \sigma^2 I)^{-1} \underline{Y},\ c_0 - \underline{k}_0^T (K + \sigma^2 I)^{-1} \underline{k}_0\right)$$

$$c_0 - \underline{k}_0^T (K + \sigma^2 I)^{-1} \underline{k}_0$$

$$\underline{k}_0 = \begin{pmatrix} k(x_1, x_0) \\ \vdots \\ k(x_n, x_0) \end{pmatrix}$$

$$c_0 = k(x_0, x_0) + \sigma^2$$

This is obtained by applying the formula of conditional distribution to

$$\begin{bmatrix} Y_0 \\ \underline{Y} \end{bmatrix} \sim N(\underline{0}, C) \qquad C = \begin{pmatrix} c_0 & \underline{k}_0^T \\ \underline{k}_0 & K + \sigma^2 I \end{pmatrix}$$

Week 8

CLASSIFICATION (Supervised learning)

The classification problem is a specific type of prediction where the response variable Y is a categorical variable, i.e. it can take values in the set $\{C_1, \dots, C_K\}$ (K classes).
(Or simply $\{1, \dots, K\}$, but remember they are just labels, there is no order).

What should we do to predict the class of a new observation Y, if we are given data: $(X_{1i}, \dots, X_{pi}, Y_i)$, $i = 1, \dots, N$?

• DECISION THEORY FOR CLASSIFICATION

We want to choose the CLASSIFIER (the predictor in this case) by minimising a LOSS FUNCTION. However, the loss functions that we have seen for parameter estimation do not make sense here, since the response variable is CATEGORICAL.

Let's say $\hat{Y} \in \{C_1, \dots, C_K\}$ is our prediction/classification, then the loss function will have a finite number of values given by all the possible combination PREDICTED CLASS - TRUE CLASS. Then, it can be written as a $K \times K$ matrix L where

L_{ij} = cost of predicting C_j when the true class is C_i

In particular $L_{ii} = 0$ (correct decision), $L_{ij} \geq 0$ if $j \neq i$

(off diagonal of matrix greater than 0 & diagonal = 0)

Simplest case (0-1 Loss): $L_{ij} = 1 \quad \forall j \neq i$

But in practice we may want to have varying costs because some mistakes are worse than others (e.g. HEALTHY PATIENT → UNHEALTHY vs. UNHEALTHY → HEALTHY)

Now we minimize the expected loss:

$$\mathbb{E}\left[L(Y, \hat{Y}(x))\right] = \mathbb{E}_X \mathbb{E}\left[L(Y, \hat{Y}(x)) \mid X\right] =$$

$$= \mathbb{E}_X \sum_{k=1}^{K} L(C_k, \hat{Y}(x)) \, P(Y = C_k \mid X)$$

C_k — $Y = C_k$

$$\Rightarrow \boxed{\hat{Y}(x) = \underset{C = C_1, \dots, C_K}{\operatorname{argmin}} \sum_{k=1}^{K} L(C_k, C_\bullet) \, P(Y = C_k \mid X = x)}$$

C_k — $Y = C_k$

If we assume the 0-1 loss function, this becomes

$$\hat{Y}(x) = \underset{C_1, \dots, C_K}{\operatorname{argmin}} \left[1 - P(Y_\bullet = C_k \mid X = x)\right]$$

$$\Updownarrow$$

$$\hat{Y}(x) = \tilde{C} \quad \text{where} \quad P(Y = \tilde{C} \mid X = x) = \max_{C_1, \dots, C_K} P(Y = C \mid X = x)$$

$\Rightarrow$ This is known as **Bayes classifier**, i.e. we classify as the most probable class according to the conditional distribution $P(Y \mid X)$.

(look pic from on 1-D input Spaces from this forums)

In any case, to minimize the expected loss function we need to estimate this conditional distribution.

HYPOTHESIS: Approaches to classification

(a) DISCRIMINATIVE APPROACH.

Define a function $f_k(x)$, called a discriminant function, and then $\hat{Y}(x) = \arg\max_k f_k(x)$

Note: $f_k(x) = P(Y = C_k \mid X)$ is a special case, sometimes called DISCRIMINATIVE MODEL

(b) GENERATIVE MODEL

Model the JOINT DISTRIBUTION $P(X, Y)$ and then classify by conditioning on X.

It is more demanding than approach (a). (Also note that X may have very large dimension)

Its advantage is that we can obtain also $P(X)$ ← marginal prob. of x, so we will be able to say if a new observation we are trying to classify has low probability under the model

⇒ OUTLIER DETECTION
outlier

9) LOGISTIC REGRESSION

Let us assume we have 2 classes C_1 and C_2.

$\Rightarrow$ $Y \sim \text{Bernoulli}(p)$

p = probability of belonging to class C_1

$$P(Y=C_k) = \begin{cases} p & \text{if } C_k = C_1 \\ 1-p & \text{if } C = C_2 \end{cases} \quad *$$

$\Rightarrow$ $Y|X \sim \text{Bernoulli}(f(x))$ (THIS IS OUR DISCRIMINATIVE MODEL)

$$p(x) = P(C_1|X,\underline{\beta}) = \frac{1}{1+\exp(-\phi(x)^T\underline{\beta})} \quad (*)$$

↳ LOGISTIC REGRESSION.

This is a special case of GENERALIZED LINEAR MODEL.

There are some good mathematical reason to choose the specific expression in (*), since it is the so-called CANONICAL LINK (but we are not going to discuss it).

In practice, the only requirement is to be a function from $\mathbb{R} \to [0,1]$, so that we can map $\phi(x)^T\underline{\beta}$ (i.e. a linear combination of our predictor) into something that is a probability.

Other options are:

PROBIT REGRESSION: $P(C_1|x,\beta) = \Phi(\phi(x)^T\beta)$

↳ C.D.F. of a STANDARD NORMAL DISTRIBUTION

COMPLEMENTARY LOG-LOG REGRESSION: $P(C_1|x,\beta) = 1 - e^{-e^{\phi(x)^T\beta}}$

It's now time we have new data $(x_{i1}, \dots, x_{ik}, y_i)$ $i = 1, \dots, N$ and we want to estimate the Logistic / logistic regression parameters β.

The likelihood is $\Rightarrow$ $P(\underline{y} \mid \beta) = \prod_{i=1}^{N} p_i^{y_i} (1-p_i)^{1-y_i}$,

(coding: $y=1$ if C_1, $y=0$ if C_2)

$\{ P(y_i = 1) = p_i$, $P(y_i = 0) = 1 - p_i$

where $p_i = \dfrac{1}{1 + \exp(-\phi(x_i)^T \beta)} = h\left(\phi(x_i)^T \beta\right)$

The log-likelihood is

$$\ell = \sum_{i=1}^{N} \left[y_i \log p_i + (1-y_i) \log (1-p_i) \right]$$

Computing the gradient: (Gradient of likelihood $\frac{\partial \ell}{\partial \beta_j}$)

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} y_i \frac{1}{p_i} \frac{\partial p_i}{\partial \beta_j} - (1-y_i) \frac{1}{1-p_i} \frac{\partial p_i}{\partial \beta_j} \quad (*)$$

Now, compute term $\frac{\partial p_i}{\partial \beta_j}$:

Now, $\dfrac{\partial p_i}{\partial \beta_j} = h'\left(\phi(x_i)^T \beta\right) \dfrac{\partial}{\partial \beta_j} \left(\phi(x_i)^T \beta\right) = h'\left(\phi(x_i)^T \beta\right) \phi_j(x_i) =$ ($\phi_j(x_i)$)

$h'(a) = + \dfrac{\exp(a)}{(1+\exp(a))^2} = \dfrac{1}{1+\exp(-a)} \left(1 - \dfrac{1}{1+\exp(-a)}\right) =$

$= h(a)\left(1 - h(a)\right)$

$$\Rightarrow \begin{cases} \dfrac{\partial p_i}{\partial \beta_j} = h\left(\phi(x_i)^T \beta\right)\left[1 - h\left(\phi(x_i)^T \beta\right)\right] \phi_j(x_i) = \\ = p_i (1-p_i) \phi_j(x_i) \end{cases} \rightarrow \text{sub this in } (*)$$

$$\Rightarrow \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{N} \left[\frac{y_i}{p_i} p_i (1-p_i) \phi_j(x_i) - (1-y_i) \frac{1}{1-p_i} p_i (1-p_i) \phi_j(x_i) \right] =$$

$$= \sum_{i=1}^{N} (y_i - p_i) \underbrace{\phi_j(x_i)}_{x_{ij}} = \left[(y - p)^T X_j \right] \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_N \end{bmatrix}$$

Gradient of likelihood

I can write this as vector inversion of my gradient as:

$$\Rightarrow \boxed{\nabla_\beta \ell = X^T (y - p)}$$

[Note: For a linear model, we get $\nabla \ell = X^T(y - X\beta)$, where $X\beta = u$, $u = H$]

(Non-linear in β & since we don't have nice linear expression of H we need hessian)

The Hessian of ℓ is then: (take derivative w.r.t β_k)

$$\frac{\partial^2 \ell(\beta)}{\partial \beta_k \partial \beta_j} = \frac{\partial}{\partial \beta_k} \frac{\partial \ell}{\partial \beta_j} = \frac{\partial}{\partial \beta_k} \left[\sum_{i=1}^{n} \phi_j(x_i)(y_i - p_i) \right] =$$

$$= \sum_{i=1}^{n} -\phi_j(x_i) \frac{\partial p_i}{\partial \beta_k} = \sum_{i=1}^{n} -\phi_j(x_i)\, p_i(1-p_i)\, \phi_k(x_i) =$$

$$= \sum_{i=1}^{n} -x_{ik}\, p_i(1-p_i)\, x_{ij} = \left(-X^T R X\right)_{kj}$$

$$R = \begin{pmatrix} p_1(1-p_1) & & 0 \\ & \ddots & \\ 0 & & p_n(1-p_n) \end{pmatrix}$$

The Hessian is negative-definite because, for $v \neq 0$,

$$\underbrace{v^T X^T}_{u^T} R \underbrace{X v}_{u} = uRu = \sum_{i=1}^{n} u_i^2 p_i(1-p_i) > 0$$

[Meaning Hessian is -vedefinate so, has one maximum.]

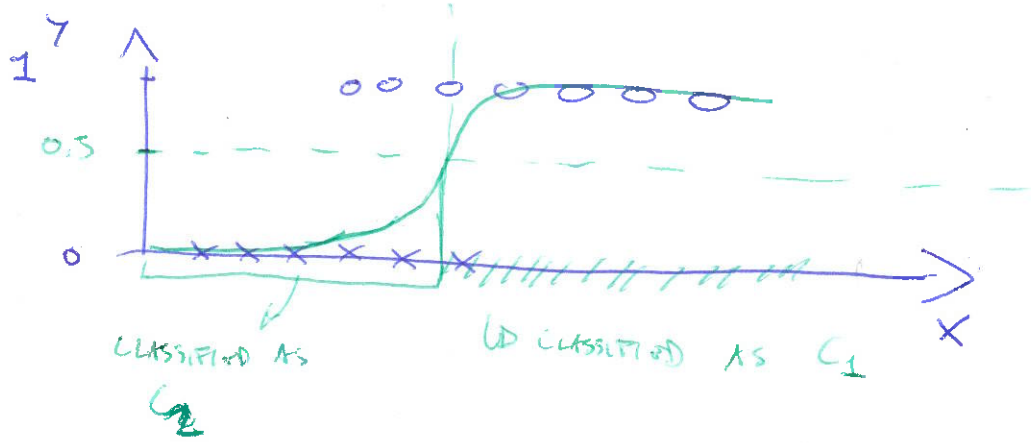
$$\Rightarrow H(\ell(\beta)) < 0$$

$\ell(\beta)$ (likelihood) has a unique maximum, so we can use NEWTON-RAPHSON METHOD to numerically solve

$$\boxed{\nabla_\beta \ell = 0}$$

In this context, it is called ITERATIVELY RE-WEIGHTED LEAST SQUARES ALGORITHM because each step is equivalent to solve a weighted least squares problem.

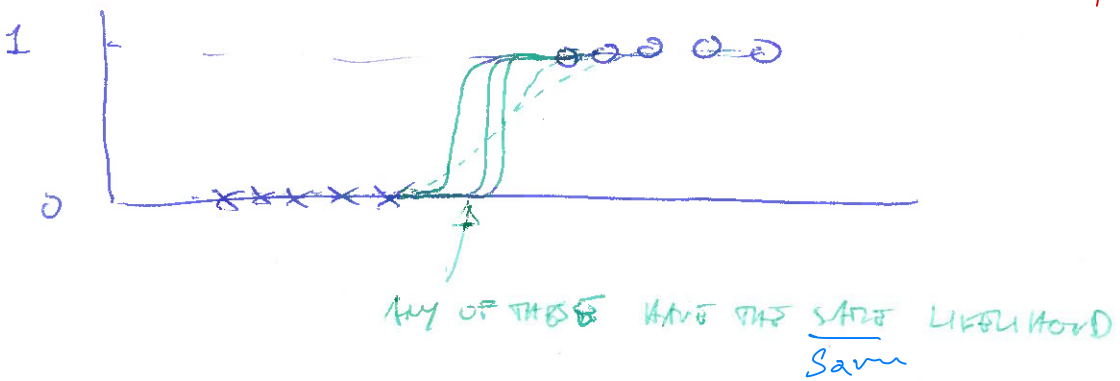
... representation : For 1D input x



The 0.5 cut-off is what is applied by the Bayes classifier if the loss function is 0-1.

Remark (see remark for page before) :

- There are identifiability issues when data are perfectly separated.



- Logistic regression can be generalized to K classes using multinomial logistic regression

$$Y \sim \text{Multinomial}(1, \pi_1, \dots, \pi_K)$$

(1 observation; π (multivariate); π_i sum to K)

$$\pi_j | x = P(Y = C_j | x) = \frac{e^{\phi(x)^T \beta_j}}{1 + \sum_{k=1}^{K-1} e^{\phi(x)^T \beta_k}}, \quad j = 1, \dots, K-1$$

$$\pi_K | x = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\phi(x)^T \beta_k}} \quad \left(\text{so that } \sum_{j=1}^{K} \pi_j = 1 \right)$$

(π_K)

BAYESIAN LOGISTIC REGRESSION:

PRIOR: $\beta \sim N(\underline{0}, \sigma_\beta^2 I)$

Posterior $\rightarrow$ $p(\beta|y) \propto p(y|\beta)\, p(\beta)$

log likelihood we written before

$$\log p(\beta|y) = \ell(\beta) - \frac{1}{2\sigma_\beta^2}\beta^T\beta + \text{const.}$$

$$\propto \underbrace{\sum_{i=1}^{N}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) - \frac{1}{2\sigma_\beta^2}\beta^T\beta}_{\hat{\ell}(\beta)}$$

this is not a Gaussian distribution.

However, it can be shown to be unimodal, so it is common practice to approximate it with a Gaussian density with the same mode. (LAPLACE approximation)

Alternatively, you can use numerical methods.

- 1-D LAPLACE APPROXIMATION: Consider a generic density

$$p(\beta) = \frac{\exp(\hat{\ell}(\beta))}{z}$$

z $\rightarrow$ NORMALIZING CONSTANT (Normalizing)

TAYLOR EXPANSION AROUND β_0 $\rightarrow$ UNIQUE MAXIMUM

first term in taylor expansion is 0 because β_0 is maximum

$\beta_0 = \max$ of my solution

$$\left\{ \hat{\ell}(\beta) \simeq \hat{\ell}(\beta_0) + \frac{1}{2}\frac{\partial^2}{\partial\beta^2}\left[\hat{\ell}(\beta)\right]_{\beta=\beta_0}(\beta-\beta_0)^2 \right\} \left(\text{since } \frac{\partial}{\partial\beta}\hat{\ell}(\beta) = 0 \right)$$

$$\Rightarrow p(\beta) \simeq \frac{e^{\hat{\ell}(\beta_0)}}{z} \exp\left(+\frac{1}{2}\frac{\partial^2}{\partial\beta^2}\left[\hat{\ell}(\beta)\right]_{\beta=\beta_0}(\beta-\beta_0)^2\right)$$

exponential of $\hat{\ell}(\beta)$

$$\Rightarrow \beta \simeq N\left(\beta_0, -\frac{1}{\frac{\partial^2}{\partial\beta^2}\left[\hat{\ell}(\beta)\right]_{\beta=\beta_0}}\right)$$

P-Dimensional Laplace Approximation:

$$\underline{\beta} \simeq N\left(\underline{\beta}_0, A^{-1}\right),$$

where $A = -\Delta \tilde{\ell}(\beta)\Big|_{\beta=\beta_0}$ ← $\tilde{\ell}$

We have $\tilde{\ell}(\beta) = \ell(\beta) - \frac{1}{2\sigma_\beta^2}\beta^T\beta$, therefore

$$-\Delta \tilde{\ell}(\beta) = X^T R X + \frac{1}{\sigma_\beta^2} I_p$$

$$R = \begin{pmatrix} \pi_1(1-\pi_1) & & 0 \\ & \ddots & \\ 0 & & \pi_N(1-\pi_N) \end{pmatrix}$$

$$\Rightarrow \boxed{\underline{\beta} \mid \underline{y} \simeq N\left(\underline{\beta}_{MAP}, S_N\right) \quad (*)}$$

where $S_N = -\Delta \tilde{\ell}(\beta)\Big|_{\beta=\beta_{MAP}}$

Therefore, we can approximate the predictive distribution:

$$p(Y=C_1 \mid \underline{x}_0, \underline{y}) = \int p(Y=C_1 \mid x_0, \beta)\, p(\beta \mid y)\, d\beta$$

$$\simeq \int h\left(\phi(x_0)^T \beta\right)\, p^{(*)}(\beta \mid y)\, d\beta$$

this needs to be computed numerically.

End of week 8

Week 3

10) Generative Models for Classification

Marginal probability → Marginal probability

Strategy:

-) Model the joint probability $p(\underline{x}, Y) = p(\underline{x} | Y)\, p(Y)$
-) Compute the conditional prob. $p(Y | X) = \dfrac{p(X|Y)\, p(Y)}{\sum_K p(X|Y=C_K)\, p(Y=C_K)}$
-) [Assuming a 1-0 loss function] Bayes classifier

$$\hat{Y}(x) = C \quad \text{where} \quad C = \underset{\text{or } C_1, \dots, C_K}{\operatorname{argmax}} \; p(Y = C_K | X = x)$$

(Alternatively, minimize the expected loss

$$\hat{Y}(x) = \underset{C \in \{C_1, \dots C_K\}}{\operatorname{argmin}} \sum_{k=1}^{K} L(C_K, C)\, p(Y = C_K | X = x)$$

-) Optionally, check $p(X = x)$ is not too small (outlier detection)

Example: 2 classes: $\{C_1, C_2\}$ — Predictor given it belongs to one of 2 classes

$$X | Y = C_K \sim N_p(\mu_K, \Sigma_K), \quad k = 1, 2$$

$$p(Y = C_K) = \pi_K \qquad (\pi_1 = \pi, \pi_2 = 1 - \pi)$$

To compare the two classes, we can simply look at the logarithm of the ratio between (Ratio between) $p(Y = C_1 | x)$ and $p(Y = C_2 | x)$:

$$\log \frac{p(Y = C_1 | x)}{p(Y = C_2 | x)} \gtrless 0 \quad \begin{cases} > 0 \Rightarrow \hat{Y}(x) = C_1 & \to \text{predicting } C_1 \text{ if } > 0 \\ < 0 \Rightarrow \hat{Y}(x) = C_2 & \to \text{" } C_2 \text{ if } < 0 \end{cases}$$

Now,

$$\log \frac{p(Y = C_2 | x)}{p(Y = C_2 | x)} = \log \frac{p(X | Y = C_1)}{p(X | Y = C_2)} + \log \frac{p(Y = C_1)}{p(Y = C_2)} =$$

($p(X | Y = C_1) \to f_1(x)$; $p(X | Y = C_2) \to f_2(x)$; $p(Y = C_1) \to \pi_1 = \pi$; $p(Y = C_2) \to \pi_2 = 1 - \pi$)

$$= \log f_1(x) - \log f_2(x) + \log \frac{\pi_1}{\pi_2} =$$

where $f_k(\cdot)$ is the p.d.f. of $N_1(\mu_k, \Sigma_k)$

$$= -\frac{1}{2}\log|\Sigma_1| + \frac{1}{2}\log|\Sigma_2| + \log\frac{\pi_1}{\pi_2} +$$

$$-\frac{1}{2}(\underline{x}-\underline{\mu}_1)^T \Sigma_1^{-1}(\underline{x}-\underline{\mu}_1) +$$

$$+\frac{1}{2}(\underline{x}-\underline{\mu}_2)^T \Sigma_2^{-1}(\underline{x}-\underline{\mu}_2) =$$

$$= \underbrace{-\frac{1}{2}\log|\Sigma_1| + \frac{1}{2}\log|\Sigma_2| + \log\frac{\pi_1}{\pi_2} - \frac{1}{2}\underline{\mu}_1^T\Sigma_1^{-1}\underline{\mu}_1 + \frac{1}{2}\underline{\mu}_2^T\Sigma_2^{-1}\underline{\mu}_2}_{\beta_0} +$$

↑ terms without x

$$+ \underbrace{\underline{\mu}_1^T\Sigma_1^{-1}\underline{x} - \underline{\mu}_2^T\Sigma_2^{-1}\underline{x}}_{+\underline{\beta}^T\underline{x}} +$$

← terms in x

$$-\frac{1}{2}\underline{x}^T\Sigma_1^{-1}\underline{x} + \frac{1}{2}\underline{x}^T\Sigma_2^{-1}\underline{x} =$$

← terms quadratic in x

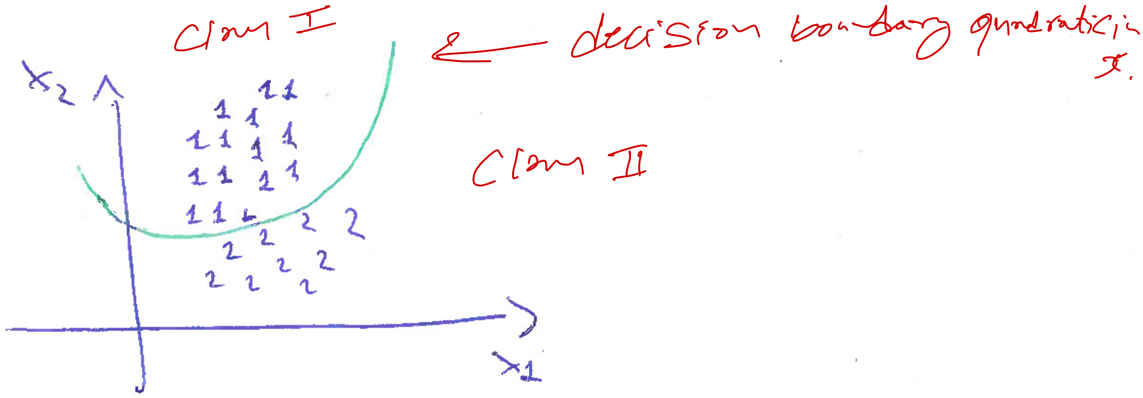
$$= \boxed{\beta_0 + \underline{\beta}^T\underline{x} + \frac{1}{2}\underline{x}^T\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\underline{x}}$$

QUADRATIC DISCRIMINANT FUNCTION

THIS METHOD IS CALLED QUADRATIC DISCRIMINANT ANALYSIS (QDA)

REMARK:
- Starting with a GENERATIVE APPROACH, we ended up with a discriminant function we could have chosen without any probabilistic model

Visualization:



• **Special case:** LINEAR DISCRIMINANT ANALYSIS (LDA) ← Special case of QDA

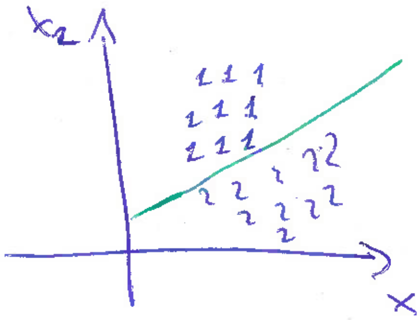
if $\Sigma_2 = \Sigma_1 = \Sigma$:

$$\log \frac{P(Y=c_1|x)}{P(Y=c_2|x)} = \underbrace{\log \frac{\pi_1}{\pi_2} - \frac{1}{2}(\underline{\mu}_1 - \underline{\mu}_2)\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)}_{\beta_0} +$$

$$+ \underbrace{(\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1} \underline{x}}_{\beta^T \underline{x}} =$$

$$= \boxed{\beta_0 + \beta^T \underline{x}}$$

LINEAR DISCRIMINANT FUNCTION



Remark: • LDA & QDA are often used also when the data are not Gaussian, simply because linear and quadratic boundaries work well as discriminative functions. However, it may make more sense to use different (data-driven) expressions for β_0 and β_1. • LDA in larger predictor space ⇒ quadratic boundary

PARAMETER ESTIMATION FOR LDA (QDA LEFT FOR EXERCISE)

(Just like Maximum likelihood estimation)

$(\underline{x}_i, y_i)$, $i = 1, \dots, N$

$$y_i = \begin{cases} 1 & \text{IF the } i\text{-th observation belongs (belongs) TO } C_1 \\ 0 & \text{ELSE} \end{cases}$$

LDA MODEL IS EQUIVALENT TO A GAUSSIAN MIXTURE WITH 2 COMPONENTS WHEN COMPLETE DATA LIKELIHOOD IS AVAILABLE:

$$p(y, x \mid \overbrace{\pi, \mu_1, \mu_2, \Sigma}^{\theta}) = \prod_{i=1}^{N} \left[\pi \, f_1(\underline{x}_i)\right]^{y_i} \left[(1-\pi) \, f_2(\underline{x}_i)\right]^{1-y_i}$$

(pdf Gaussian)

$$\log p(y, x \mid \theta) \propto \sum_{i=1}^{N} y_i \left(\log \pi - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\underline{x}_i - \underline{\mu}_1)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_1) \right) + (1 - y_i) \left(\log (1-\pi) - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (\underline{x}_i - \underline{\mu}_2)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_2) \right)$$

max. w.r.t π first

$$\downarrow$$

$$\max_{\pi} \log p(y, x \mid \theta) = \max_{\pi} \underbrace{\sum_{i=1}^{N} y_i \log \pi + (1 - y_i) \log (1-\pi)}_{(*)} =$$

$$N_1 = \sum_{i=1}^{N} y_i \quad \rightarrow \max_{\pi} \; N_1 \log \pi + (N - N_1) \log (1-\pi)$$

$$\frac{\partial (*)}{\partial \pi} = \frac{N_1}{\pi} - \frac{(N - N_1)}{1-\pi} = 0$$

$$\Rightarrow \frac{\pi}{1-\pi} = \frac{N_1}{N - N_1} \Rightarrow \boxed{\hat{\pi} = \frac{N_1}{N}}$$

Taking derivative set it equal to zero as in MLE.

Now maximize w.r.t μ_1:

$$\cdot) \max_{\mu_1} \log\left(p(y,x|\theta)\right) = \max_{\mu_1} \sum_{i=1}^{N} y_i \left[-\frac{1}{2}(\underline{x}_i - \underline{\mu}_1)^T \Sigma^{-1} (\underline{x}_i - \underline{\mu}_1)\right] =$$

$$= \max_{\mu_1} \sum_{j=1}^{N_1} -\frac{1}{2}(\underline{x}_j^{(1)} - \underline{\mu}_1)^T \Sigma^{-1} (\underline{x}_j^{(1)} - \underline{\mu}_1)$$

← Sum over only one for which $y_i = 1$

where $\underline{x}_j^{(1)}$ is the j-th obs for which $y_i = 1$

$$\Rightarrow \text{FROM DERIVATIVE NORMAL RESULTS}: \quad \boxed{\hat{\underline{\mu}}_1 = \frac{1}{N_1} \sum_{j=1}^{N_1} \underline{x}_j^{(1)}}$$

← We have done this for Gaussian Multivariate

ANALOGOUSLY: $\boxed{\hat{\underline{\mu}}_2 = \frac{1}{N_2} \sum_{j=1}^{N_2} \underline{x}_j^{(2)} = \frac{1}{N - N_1} \sum_{j=1}^{N-N_1} \underline{x}_j^{(2)}}$

Maximize w.r.t. Σ:

$$\cdot) \max_{\Sigma} \log\left(p(y,x|\theta)\right) =$$

$$= \max_{\Sigma} \sum_{i=1}^{N} \left[-\frac{1}{2}\log|\Sigma| - \frac{y_i}{2}(\underline{x}_i - \hat{\underline{\mu}}_1)^T \Sigma^{-1} (\underline{x}_i - \hat{\mu}_1) + \right.$$

$$\left. + -\frac{1-y_i}{2}(\underline{x}_i - \hat{\mu}_2)^T \Sigma^{-1} (\underline{x}_i - \hat{\mu}_2)\right] =$$

$$= \max_{\Sigma} -\frac{N}{2}\log|\Sigma| - \frac{1}{2}\text{trace}\left(\Sigma^{-1} \sum_{j=1}^{N_1} (\underline{x}_j^{(1)} - \hat{\mu}_1)(\underline{x}_j^{(1)} - \hat{\mu}_1)^T\right) +$$

$$- \frac{1}{2}\text{trace}\left(\Sigma^{-1} \sum_{j=1}^{N_2 - N_1} (\underline{x}_j^{(2)} - \hat{\mu}_2)(\underline{x}_j^{(2)} - \hat{\mu}_2)^T\right) =$$

$$= \max_{\Sigma} -\frac{N}{2}\log|\Sigma| - \frac{N}{2}\text{trace}\left(\Sigma^{-1} V\right)$$

where $V = \frac{1}{N}\left[\sum_{j=1}^{N_1} (x_j^{(1)} - \hat{\mu}_1)^T (x_j^{(1)} - \hat{\mu}_1)^T + \sum_{j=1}^{N-N_1} (x_j^{(2)} - \hat{\mu}_2)(x_j^{(1)} - \hat{\mu}_2)^T\right]$

from multivariate Normal results
MLE multivariate normal parameters:

$$\Rightarrow \boxed{\hat{\Sigma} = V}$$

Remark: the unbiased estimator

$$\hat{\Sigma} = \frac{1}{N-2}\left(\sum_{j=1}^{N_1} (x_j^{(1)} - \hat{\mu}_1)(x_j^{(1)} - \hat{\mu}_1)^T + \sum_{j=1}^{N-N_1} (x_j^{(2)} - \hat{\mu}_2)(x_j^{(2)} - \hat{\mu}_2)^T \right)$$

is sometimes preferred.
Sometimes preferred

- Further remark on this written after Page (Multiclass LDA/QDA).

MULTI-CLASS LDA / QDA

$$X|Y=C_k \sim N(\mu_k, \Sigma_k) \quad , k=1,\dots,K$$

We describe the decision rules we obtain for any 2 pairs of classes as a discriminative function; $\underset{\max}{\arg} \log P(Y=C_k|X)$:

QDA: $$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log\pi_k$$

LDA: $$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$$

$x^T \to v_{gc}^T$

then we classify by maximizing the discriminative functions:

$$\hat{Y}(x) = \underset{k=1,\dots,K}{\arg\max}\ f_k(x)$$

$D_k(x)$ → Discriminative function

- Remark

REMARK : • Both LDA and QDA require $\hat{\Sigma}^{-1}$.

p = no. of predictors (dimensionality)
N = no. of observations

Therefore we need $\hat{\Sigma}$ to be invertible, i.e.

$$\text{Rank}(\hat{\Sigma}) = p \Rightarrow N > p$$

Therefore LDA and QDA do not work in HIGH-DIMENSIONAL settings where $p > N$, unless some dimensional reduction is performed first.

• To estimate parameters, you need a genuine sample from (X, Y) (joint dist. of X & Y), it does not work if we select X values as in linear regression. (not work)

- Other approaches to Generative Classification : (Non-Examinable Below)

:) **Regularized Discriminant Analysis**

the covariance of the 2 groups is assumed to be different, but the estimates are regularised so that they are not too different:

$$\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

($\hat{\Sigma}_k$: QDA estimate; $\hat{\Sigma}$: LDA estimate)

(In between LDA & QDA.)

α is chosen by cross-validation.

:) **Kernel Density Classification**

the density $f_k(x) = P(X \mid Y = C_k)$ is estimated from the data using Kernel Density Estimation:

$$\hat{f}_k(x) = \frac{1}{N_k} \sum_{j=1}^{N_k} \frac{k_\lambda(x, x_j^{(k)})}{\lambda^p}$$

where $k_\lambda(\cdot,\cdot)$ is a kernel function, for example

the Gaussian kernel: $k_\lambda(x,y) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\frac{\|x-y\|^2}{\lambda^2}\right)$

then, $$P(Y = C_k \mid X = x_0) = \frac{\hat{\pi}_k \hat{f}_k(x_0)}{\sum_{j=1}^{J} \hat{\pi}_j \hat{f}_j(x_0)}$$

Naive Bayes Classifier

Since ~~Bayes~~ density estimation does not work well in high dimension, (see the curse of dimensionality)

a popular alternative is to approximate the multivariate density with:

$$f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j)$$

(i.e., assuming independence of the predictors)

this works well because it drastically decreases the variance of the estimates, even if we are including some bias due to the incorrect approximation)

incorrect

bias

End of week 9

Week 10

11) Separating Hyperplanes and SVM

We have seen in the past lectures different methods to estimate linear boundaries between decision regions, based on probabilistic models either for (X, Y) or $Y|X$.
In alternative approach is to ~~directly~~ learn the "best" linear decision boundaries from the data, without assuming a model for our observations. For simplicity, we restrict (restrict) our discussion to the 2-classes case.

• Separating hyperplanes

A separating hyperplane is a linear decision boundary $\underline{\phi}(\underline{x})^T \underline{\beta}$, such that

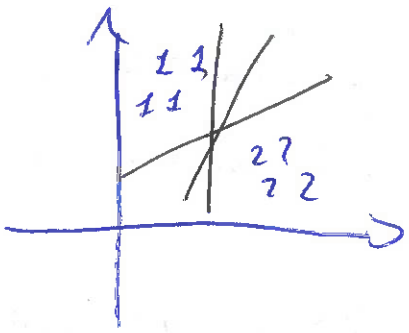
$$\underline{\phi}(\underline{x}_i)^T \underline{\beta} \begin{cases} > 0 & \text{if } Y_i = C_1 \\ < 0 & \text{if } Y_i = C_2 \end{cases}$$

The classifier is then $y(x_0) = \text{sign}\left(\phi(x_0)^T \beta\right)$.

Note that: a). DATA MAY NOT BE SEPARABLE FOR THE CHOSEN FEATURE (PREDICTOR) SPACE.

b). IF DATA ARE SEPARABLE, THE SEPARATING HYPERPLANE IS NOT UNIQUE

problem



Solution to a) may be choosing a different set of predictors, or using SVM.

Solution to b) is selecting the OPTIMAL SEPARATING HYPERPLANE.

OPTIMAL SEPARATING HYPERPLANES :

$$y_i = \begin{cases} 1 & \text{if } i \in C_1 \\ -1 & \text{if } i \in C_2 \end{cases}$$

Coding

Criterion we want to maximize

$$\begin{cases} \hat{\beta} = \underset{\beta}{\operatorname{argmax}}\ M \\ \|\beta\| = 1 \\ y_i \cdot \phi(x_i)^T \beta \geq M, \quad i = 1, \dots, N \end{cases}$$

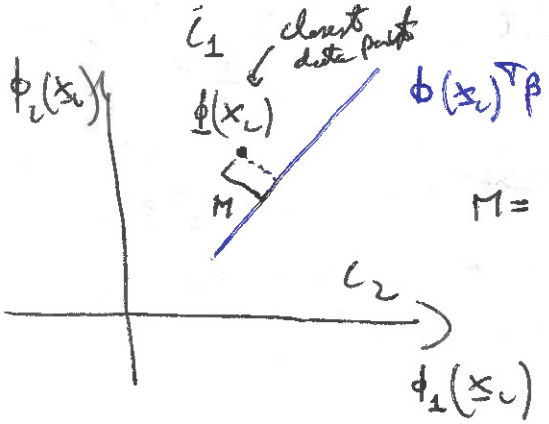
← Same number called Margin

Thus set of conditions imply that all points are at distance at least M from the decision boundary.

M is called the MARGIN.

($\|\beta\| = 1$ is required because the hyperplane does not change if we rescale the coefficients, i.e. $\phi(x_i)^T \beta = 0 \Leftrightarrow \phi(x_i)^T \lambda\beta = 0$ $\lambda \neq 0$)

2D example



$M = \phi(x_i)^T \beta$ is the smallest of this is largest possible one for each data points

- $y_i \cdot \phi(x_i)^T \beta \geq M$ guarantees that the distance of the closest point from the boundary (on either side) is larger or equal to M
($\phi(x_i)^T \beta \leq M$ if i-th obs comes from C_2)

- We can get rid of the $\|\beta\| = 1$ constraint by rescaling:

$$\frac{1}{\|\beta\|} \left(y_i \cdot \phi(x_i)^T \beta \right) \geq M \quad \text{or} \quad y_i \cdot \phi(x_i)^T \beta \geq M \|\beta\|$$

and then arbitrarily choose the norm of β, $\|\beta\| = \frac{1}{M}$

We get the equivalent problem:

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\beta\|^2$$
$$y_i[\phi(x_i)^T\beta] \geq 1, \quad i = 1..,N$$

(*) becomes minimizing problem

Solving problem (*):

It is a constrained optimization problem, so we use Lagrange multipliers:

+ KKT conditions

(PRIMAL PROBLEM) $$L = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i \phi(x_i)^T \beta - 1\right]$$

Lagrangian

Lagrange Multipliers $\alpha_i \geq 0$

Note: we want to penalize NEGATIVE VALUES OF THE PENALTY TERMS

$$\frac{\partial L}{\partial \beta_j} = 0 \Rightarrow \boxed{\beta_j = \sum_{i=1}^{N} \alpha_i y_i \phi_j(x_i)}$$

this leads to the Wolfe DUAL problem

$$(\max) \quad L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k \phi(x_i)^T \phi(x_k)$$

Subject to $\alpha_i \geq 0$, $\sum_{i=1}^{N} \alpha_i y_i = 0$, $i = 1, \ldots, N$

Lagrangian dual

Is easier to solve numerically

- **KKT CONDITIONS** :
(KARUSH-KUHN-TUCKER)

Gradient of Lagrange

- $\nabla L = 0 \Rightarrow \quad \underline{\beta} - \sum_{i=1}^{N} \alpha_i y_i \phi(\underline{x}) = 0$

- $y_i \phi(\underline{x}_i)^T \underline{\beta} - 1 \geq 0$ ← constraint

- $\alpha_i \left[y_i \phi(\underline{x}_i)^T \beta - 1 \right] \geq 0 \qquad (\#)$

$(\alpha_i > 0)$ → lagrange multipliers

Condition (#) implies that either $\alpha_i = 0$ OR $\boxed{y_i \phi(\underline{x}_i)^T \beta = 1}$

↳ this means the i-th obs is one of the closest ones to the boundary, since its distance is equal to the MARGIN

Therefore: $y_i \phi(\underline{x}_i)^T \underline{\beta} = 1$

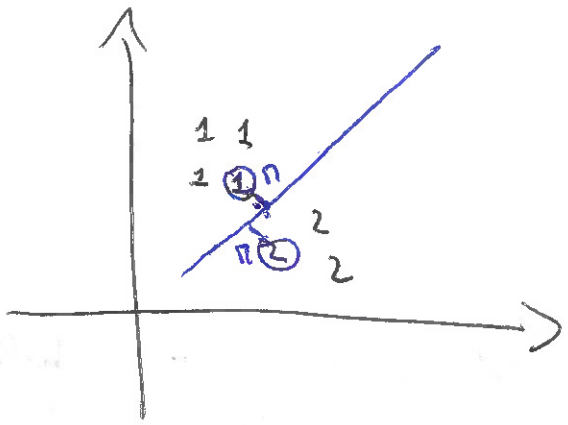
If $\alpha_i > 0$, the i-th obs. (it lies on the MARGIN) is close # to the boundary and it impacts the expression of the optimal hyperplane → x_i is called **SUPPORT POINT**

If $y_i(\phi(\underline{x}_i)^T \beta) > 1$, the observation is further away from the boundary, $\alpha_i = 0 \Rightarrow$ this observation does not impact the expression of the boundary $\left(\text{since } \vec{\beta} = \sum_{i=1}^{N} \alpha_i y_i \phi(\underline{x}_i) \right)$

REMARK: • Since the dual problem depends on $\phi(x_i)^T \phi(x_j) = k(x_i, x_j)$ and only a small fraction of the data points determines the solution, this approach is referred to as a

SPARSE KERNEL METHODS.

2D example:



• the classifier is $\hat{Y}(x) = \text{sign}\left(\phi(x)^T \hat{\beta}\right)$

classifier

• **SUPPORT VECTOR MACHINES** (FOR CLASSIFICATION)

the optimal hyperplane is a special ~~case~~ example of support vector CLASSIFIER.

It is so called because the classification rule depends on a few SUPPORT VECTORS (i.e. the ones lying on the MARGIN.)

The general formulation allows for the classes to overlap (a little).

So now we want to maximize M, but allowing some points to be on the "wrong" side of the boundary.

We therefore reformulate the optimization problem in the following way:

$$
\begin{cases}
\max_{\beta,\ \|\beta\|=1} M \\
y_i\left(\phi(x_i)^T\beta\right) \geq M\left(1-\varepsilon_i\right) \quad \forall i \\
\varepsilon_i \geq 0 \\
\sum_{i=1}^{n} \varepsilon_i \leq C \leftarrow \text{additional parameter}
\end{cases}
$$

this means we are allowing data points to be a little bit closer than M to the margin, as long as the sum of the slack variables ε_i is not too large.

As before, we can rewrite this problem as:

$$
\begin{cases}
\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\|\beta\|^2 \\
y_i\left(\phi(x_i)^T\beta\right) \geq 1-\varepsilon_i \\
\varepsilon_i \geq 0 \\
\sum_{i=1}^{n} \varepsilon_i \leq C
\end{cases}
$$

• Solving the SVM problem :

We can write the Lagrangian:

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i(\phi(x_i)^T\beta) - (1-\xi_i)\right] - \sum_{i=1}^{N}\lambda_i\xi_i$$

$\sum_{i=1}^{N}\xi_i \leq C$ $\qquad$ $y_i(\phi(x)^T\beta) \geq (1-\xi)$ $\qquad$ $\xi_i > 0$

KKT conditions :

- $\nabla L_p = 0 \Rightarrow \hat{\beta} = \sum_{i=1}^{N}\alpha_i y_i \phi(x_i)$
- $\alpha_i = C - \lambda_i$ $\qquad$ $\alpha_i\left[y_i\phi(x_i)^T\beta - (1-\xi_i)\right] = 0$
- $\lambda_i\xi_i = 0$ $\qquad$ $y_i\phi(x_i)^T\beta - (1-\xi_i) \geq 0$

We get that $\alpha_i > 0 \Leftrightarrow y_i\phi(x_i)^T\beta = 1 - \xi_i$

IF $\hat{\xi}_i = 0$ and $y_i\phi(x_i)^T\beta = 1 \Rightarrow 0 < \alpha < C$

IF $\hat{\xi}_i > 0$ and $y_i\phi(x_i)^T\beta = 1 - \hat{\xi}_i \Rightarrow \alpha = C$

the problem needs to be solved numerically (usually through the simpler dual problem)

Remarks : • SVM can be generalized with different kernels:

$$k(\underline{x}_i, \underline{x}_j) = \phi(\underline{x}_i)^T \phi(\underline{x}_j)$$

$$k(\underline{x}_i, \underline{x}_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right) \quad \text{GAUSSIAN}$$

$$k(x_i, x_j) = (1 + x^T x)^d \quad \text{POLYNOMIAL}$$

and the classifier rewritten as

$$\hat{Y}(\underline{x}) = \text{sign}\left(\phi^T(\underline{x})\beta\right) = \text{sign}\left(\phi^T(\underline{x})\left(\sum_{i=1}^{N} \alpha_i y_i \phi(\underline{x}_i)\right)\right) =$$

$$= \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i \phi^T(\underline{x}_i)\phi(\underline{x})\right) =$$

$$= \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i k(\underline{x}, \underline{x}_i)\right)$$

- the constant C is playing the role of a regularization parameter and needs to be chosen (for example by cross-validation)

- the constant C depends on how large your model is (either in terms of feature space $\phi_1(\underline{x}) \dots \phi_p(\underline{x})$ or complexity of the kernel)

- Some people make a distinction between SOFT SVM and HARD SVM (i.e. separating hyperplanes)

End of week 10