

L'OGGETTIVO DI UN GENERICO PROBLEMA DI APPRENDIMENTO
È QUELLO DI RIUSCIRE A PARTECIPARE I PARAMETRI OTTIMI W DI UN
CERTO SPAZIO DELLE IPOTESI PARAMETRIZZATE A

CONSIDERANDO LA CLASSE DEI PROBLEMI DI REGRESSIONE
UNIVERSITÀ SI HAVO COPIE D'OSSERVAZIONI DEL TIPO:

$$(x_i, t_i)$$

Dove

- x_i sono le osservazioni della variabile INDEPENDENT
- t_i è la variabile INDEPENDENT

Si ha quindi che la variabile INDEPENDENT è CHIAMATA
A PUGLI INDEPENDENT TANTO CHE UNA FUNZIONE f
INCognita E UNCARE NEI SUOI PARAMETRI W

→ LA FUNZIONE CHE LEGA LE VARIABILI INDEPENDENT A QUESTE INDEPENDENT
TANTO CHE COMBINAZIONE LINEARE DI PARAMETRI SONO SOLUZIONI

$$y(x|w) = w_0 + w_1 x = \underline{w}^T \begin{bmatrix} 1 \\ x \end{bmatrix}$$

NELLA MIGLIOR APPROXIMAZIONE SI DICE CALCOLATA

LA LOSS L'È LA PARTE MIGRA L'ERRORE DEL
PROGETTO W' CHE ABBIANO NEL NOSTRO

L'È PENSATO CON RIFERIMENTO AI PARAMETRI, SIA
DEL PROGETTO

$$D = \{(x_i, t_i) \mid i=1, \dots, N\}$$

→ è la somma degli errori quadratici tra le funzioni incognite
approssimate ed i valori obiettivi

$$L(\underline{w} | D) = \frac{1}{2} \sum_{i=1}^N \left[y(x_i | \underline{w}) - t_i \right]^2$$

↓
errore quadratico

valore
TARGET

valore incognito

Dove $y(x_i | w) = w_0 + w_1 x_i$

QUESTO APPROCCIO TUTTI NOI È PIÙ OFFICIALE QUANDO
SI SONO DEI PUNTI **DISTORZI** CHE FENOMENO AD
AUMENTARE L'ERRORE QUADRATICO

POSSIAMO DI UN APPROCCIO OBBLIGATORIO SO CHE
PROBABILISTICO, SI HA UN CIE, PER IL RICORSIONE
LINEARE, DOVETE CALCOLARE IL **STIMA DI MASSIMA**

VEROSIMILANZA (ML) DAI PARAMETRI OTTIMI W

PRENDENDO DAL DATASET D

IL MODELLO CHE CI PIÙ SERVIRÀ PER LA REGRESSIONE UTILIZZANDO LE COMBINAZIONI TRA LE DUE VARIABILI

INPUT:

$$\underline{x} = (x_1, \dots, x_D)^T$$

→ DATA IN VETTORE
 IN R^m MA PIÙ
 IN R^D

OTTERENDO QUINDI:

$$y(\underline{x} | \underline{w}) = w_0 + w_1 x_1 + \dots + w_D x_D$$

↪ È UNA FUNZIONE LINEARE NEI PARAMETRI w
NEI VALORI IN MEZZO

SE w_0 È UNO NUMERO POSITIVO CHE FA FUNZIONARE UNO NELLA PIANURA, DIVENTA LINEARE QUANDO LA PIANURA È LINEARE NEI VALORI DI INCEDENZA. SE UNO QUINDI AD AUMENTARLE LA DIMENSIONE POSSIAMO DI UTILIZZARE COMBINAZIONI LINEARI DI FUNZIONI DI LINEA FISSATE

DALL'INPUT:

$$y(\underline{x}, \underline{w}) = w_0 + \sum_{j=1}^{m+1} w_j \phi_j(\underline{x})$$

→ FUNZIONE BASE

DOVE w_0 È IL BIAS, ANDRA IN POSSIBILE OFFSET

nel caso, mentre ψ_j è la **funzione base**. Ma prendendo in input i valori x e restituendo una scena.

Alcuni esempi:

$$\phi_i(x) = x^i \quad \phi_i = \exp\left\{-\frac{(x-\mu_i)^2}{2s^2}\right\} \quad \phi_i(x) = \tanh(x)$$

Polynomial Gaussian Sigmoid

Dati che vogliamo utilizzare un approccio probabilistico vediamo che è utile di fare così

$$t = y(\underline{x}, \underline{w}) + \varepsilon$$

È ovviamente una funzione ipotesi più il **RUMORE** **Gaussiano a mezzanotte**. Possiamo quindi scrivere:

→ Probabilità di avere un valore t DATI i valori in ingresso \underline{x} ED i parametri vicari \underline{w} sulla base della presenza di rumore β

$$p(t | \underline{x}, \underline{w}, \beta) = N(t | y(\underline{x}, \underline{w}), \beta^{-1})$$

Parametri del rumore

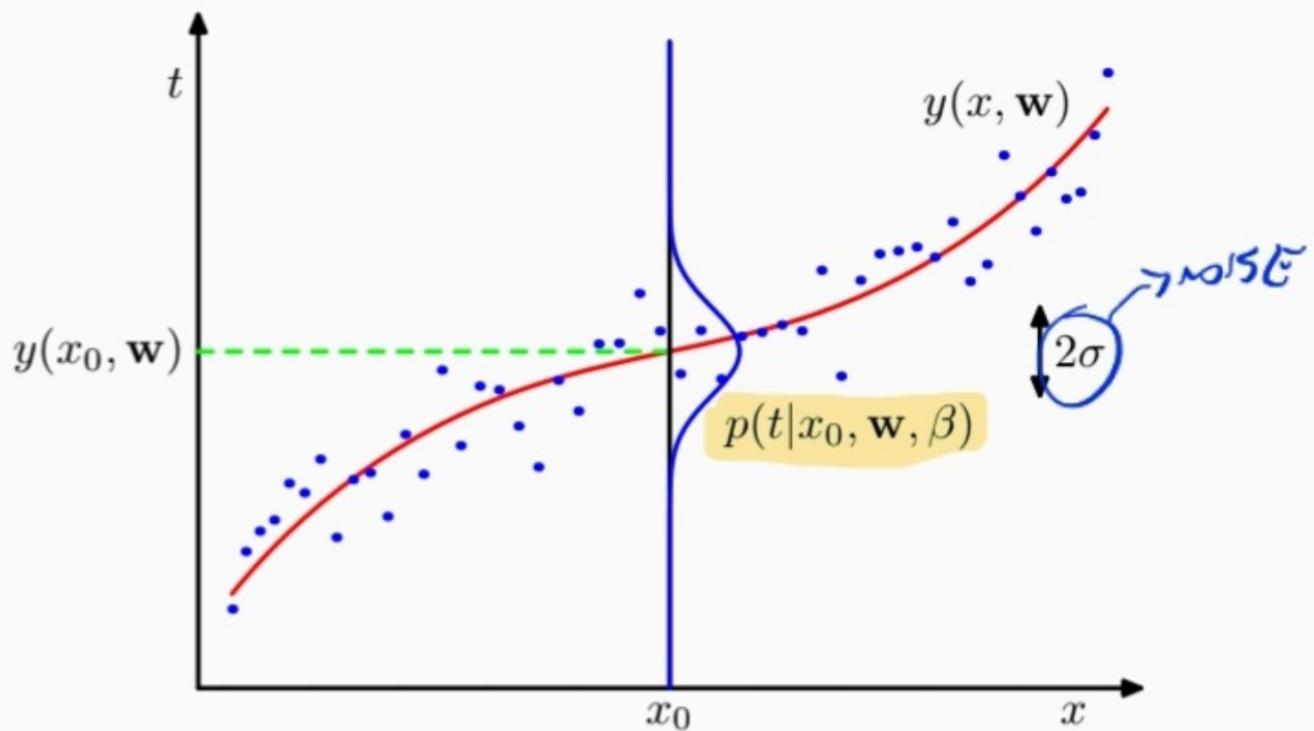
Funzione ipotesi

$$\beta^{-1} = \frac{1}{\sigma^2}$$

Allora possiamo ottenere una **distribuzione di probabilità** che lega il target ai valori in ingresso e nel frattempo rispetto ad un certo valore di noise.

Punto, per passare allo **probabilità**, il predittore **gaussia** così la **previsione** risulta con una **gaussiana**.

CENTRALS NGL VERS



DATI PENS IL DATO DI INPUT

$$\underline{x}_s \{x_1 - x_n\}$$

E I CORRESPONDENTI VALORI TAZZO $\underline{t} = \{t_1, \dots, t_n\}^T$

LA LINEARE INSIEME LA PROBABILITA' CORRESPONDENTI DEI DATI OSSERVATI

POSSIAMO SCRIVERE LA **LINELHOOD** SOTTO GLI SPECIFICI MODELLI COME:

$$P(\underline{t} | \underline{x}, \underline{w}, \beta) = \prod_{n=1}^N N(t_n | y(\underline{x}_n, \underline{w}), \beta^{-1})$$

$$= \prod_{n=1}^N N(t_n | \underline{w}^T \phi(\underline{x}_n), \beta^{-1})$$

\hookrightarrow SO' LA LINELHOOD!

ONE OF POSSIBLE SEMIPLICATIVE APPROXIMATIONS IS
LINEARIZING:

$$\frac{1}{\sqrt{2\pi}\beta} \exp\left\{-\frac{\beta}{2}(t - w^T \phi(x_m))^2\right\}$$

$$\ln p(t | \underline{w}, \beta) = \sum_{m=1}^N \ln N(t_m | \underline{w}^T \phi(\underline{x}_m), \beta^{-1})$$
$$= \underbrace{\frac{N}{2} (\ln \beta - \ln(2\pi))}_{\text{CONSTANT}} - \underbrace{\frac{\beta}{2} \sum_{m=1}^N (t_m - \underline{w}^T \phi(\underline{x}_m))^2}_{\text{TERM PROPORTIONAL}}$$

ADESSO VOGLIAMO MASSIMIZZARE IL LIQUIDAZIONE PER \underline{w} .
SI CALCOLA QUINDI IL GRADIENTE:

$$\nabla \ln p(t | \underline{w}, \beta) = \beta \sum_{m=1}^N (t_m - \underline{w}^T \phi(\underline{x}_m)) \phi(\underline{x}_m)^T$$

E LO SI UGUALE A ZERO:

$$\sum_{m=1}^N t_m \phi(\underline{x}_m)^T - \underline{w}^T \left(\sum_{m=1}^N \phi(\underline{x}_m) \phi(\underline{x}_m)^T \right) = 0$$

E SOLVENDO PER \underline{w} :

$$\underline{w}_{M2} = (\underline{\phi}^T \underline{\phi})^{-1} \underline{\phi}^T \underline{t}$$

\rightarrow w MAX LIQUIDAZIONE

\hookrightarrow w CHE MASSIMIZZA IL LIQUIDAZIONE

QVI?

TUTTE LE FUNZIONI BASE

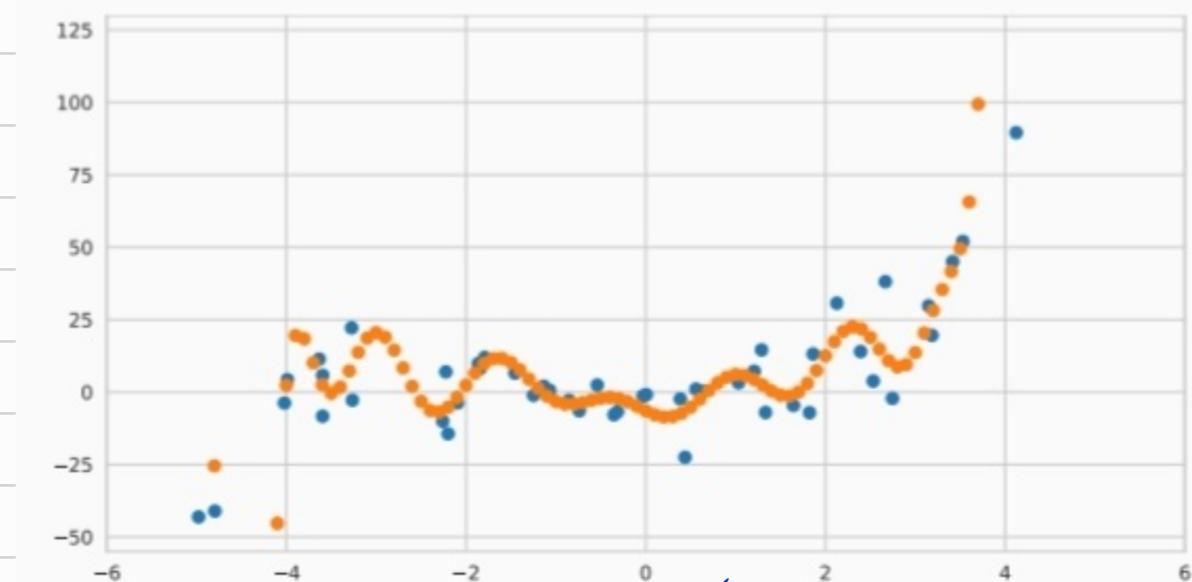
$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N)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

Dove:

- Usa **NOX** contiene **TUTTE LE Funzioni Base**
DGL minima Sample corrispondente (dimensione **M**)
- Usa **colori** Gli corrispondenti **Funzioni Base**
corrispondenti per i training samples (dimensione **N**)

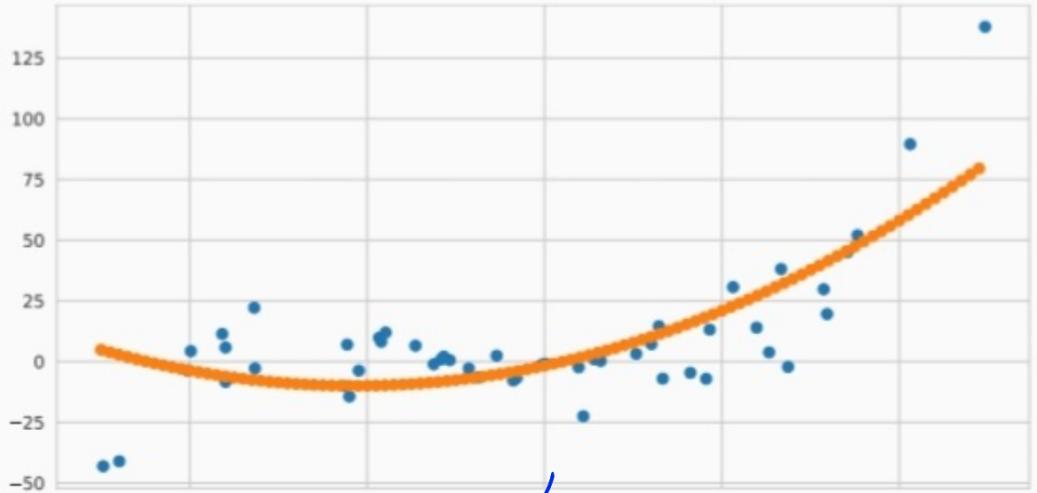
PER TUTTI I PUNTI **M** Gli Grandi SI HA **OVERFITTING**

→ quando si usa una funzione base troppo complessa!



→ polinomio di grado 20

EM TRAPS Preciso Otimizar CUSTOS ITINERARIA



ANALOGO AS ISPECIACOES DA MENUTDAD DE PESO
 SISTEMA DEVE MAX UNIFORME OTIMIZAR CUSTOS, VELAS
 SÓRIO GRADO DE DISPARO PROVAR QUANDO IL CUSTO DEL
 PESO É ACCESIVO.

PUNTO MÉTODOS CHEM IN PROBLEMS OF REGRESSION
 ALGORITMO PARA OSAS MINIMIZE A LOSS NO ANTES SE
REGULARIZAÇÃO:

$$E_0(\underline{w}) + \lambda E_w(\underline{w})$$

Loss
Regularização

DESTE PONTO PODE USAR \hat{g} IL COSTO DELL' (L_2)

$$E_w(\underline{w}) = \frac{1}{2} \underline{w}^T \underline{w}$$

SE \hat{g} NAO IL
 VALORE DI λ SI DI
 PODE PESO NO REGULARIZZAZ
 DIRE MELHORAR A PREVISÃO

La complessità = dimensione
e quindi si riduce
overfitting

Perché la funzione di lavoro totale di minimizzazione è
data da?

$$E(w) = \frac{1}{2} \sum_{n=1}^N (t_n - w^T \phi(w))^2 + \frac{\lambda}{2} w^T w$$

REGRESSIONE LINEARE BAYESIANA (introd)

Fino ad ora abbiamo visto che utilizzando la **metodica**
lineare oppure **least squares**, più potrete le feature
di overfitting nel caso in cui modelli complessi siano
utilizzati utilizzando dati-set limitati. Tuttavia, andiamo a
limitare il grado del polinomio utilizzando c'è un effetto
opposto, quindi si va a limitare la flessibilità del modello
nel curvare le relazioni intriseche presenti.

Tramite l'introduzione del **regularizzazione** è possibile
limitare gli effetti dell'overfitting, ma però c'è un valore
ottimale dei parametri λ ?

Nel caso quindi della **regressione lineare Bayesiana** si va
a incorporare il concetto di probabilità Bayesiana per
stima i parametri del modello. Tramite questo tipo di approccio

È POSSIBILE MODELLARE CON GLI SDI IL VERSO PREVISTO DELLE VARIABILI DIPENDENTI, MA ANCHE LA **DISTRIBUZIONE DEI PARAMETRI DEL MODELLO**.

BIAS-VARIANZA DECOMPOSITION

QUESTA DECOMPOSIZIONE DI **ERRORE COMBOSO (BIAS)** E **VARIANZA** AIUTA A COMPRENDERE I FATTORI CHE INFLUENZANO LE PREVISIONI COMPLESSIVE DI UN MODELLO, IN MODO DA POTERLO MIGLIORARE.

• BIAS (\rightarrow OVERFITTING)

INDICA L'ERRORE SISTEMATICO DEL MODELLO. INDICA QUANTO IL MODELLO APPROSSIMA MOLTE I DATI DELL'ADDESTRAMENTO. QUINDI UN MODELLO CON BIAS ELEVATO TENDE AD ESSERE TROPPO SEMPUCE E NON HA QUINDI A CATENA DI COMPLESSITÀ E LE RELAZIONI PRESENTE NELI DATI

• VARIANZA (\rightarrow UNDERFITTING)

LA VARIANZA MISURA IL VARIABILITÀ DELLE PREVISIONI DEL MODELLO SU DATI DIVERSI. QUINDI UN MODELLO HA UNA VARIANZA Eccessivamente COMPLESSO E QUINDI SI ADATTA TROPPI BENI AI DATI DI ADDESTRAMENTO, MA GENERALIZZA MOLTE AL DATI CHE CI SONO SCARICHI. LA VARIANZA MINIMA QUINDI QUINDI IL MODELLO È TROPPO COMPLESSO \rightarrow QUINDI SI HA OVERFITTING!

QUINDI LA DECOMPOSIZIONE BIAS-VARIANZA È SCRITA COME:

$$\text{ERRORE TOTALE} = \text{BIAS}^2 + \text{VARIANZA} + \text{RUMORE}$$

\hookrightarrow EXPECTED LOSS

L'OBETTIVO È QUINDI RICAVARE UN MODELLO CHE MINIMIZZI L'ERRORE

TOTALE PENSANDO AI DUE COMPONENTI CHE LO COSTITUISCONO, SI RICADE QUINDI IN UN TRADE-OFF:

- AUMENTANDO LA COMPLESSITÀ DEL MODELLO (AD ESEMPIO AUMENTANDO IL NUMERO DI PARAMETRI) SI VA A RIDURRE IL BIAS, POICHÉ IL MODELLO PUÒ ADATTARSI MIGLIORAI AL DATI
- TUTTAVIA, SE SI AUMENTA LA COMPLESSITÀ, VOGLIO AUMENTARE ANCHE LA VARIANZA

\Rightarrow SI CERCA UN EQUILIBRIO IN MODO CHE IL MODELLO SIA IN grado di CATTURARE LA STRUTTURA DEI DATI SENZA PERDERE DIVENTARE TROPPO COMPLESSO DA PERDERE IL CAPACITÀ DI GENERALIZZARE SUL MOLTI DATI.

ALCUNI METODI PER AFFRONTARE IL TRADE-OFF SONO LE REGULARIZZAZIONI E LA CROSS-VALIDAZIONE.

NECESSITA UNA BAYESIANA

IL PROBLEMA DELL'APPROCCIO DESCRITO DALLA MAXIMUM LIKELIHOOD È IL FATTO CHE NON CI PERMETTE DI CAPIRE RENDIMENTO QUANTO EFFETTUAMENTE SIA PROBABILE LA SOLUZIONE CHE ABBIAMO TROVATO, OVVERO QUANTO "CREDO" CHE QUESTA SOLUZIONE SIA

ESTIMAZIONE AFFIDABILE.

L'OGGETTO è quindi quello di ricavare una **PIA** al valore stimato. Ricorrendo al **TEOREMA DI BAYES**:

$$P(w|D) = \frac{\text{DATA LIKELIHOOD} \times \text{PRIOR}}{\text{EVIDENCE}}$$

$$= \frac{P(D|w) P(w)}{P(D)}$$

E' la likelihood classica discussa prima

SUPPIAMO CHE LA LIKELIHOOD HA LA SEGUENTE FORMA:

$$P(\bar{e}|\bar{w}, \beta) = \prod_{i=1}^N N(e_i | \bar{w}^T \phi(x_i), \beta^{-1})$$

E abbiamo quindi bisogno di trovare una **DISTRIBUZIONE** a priori che esprima il **PRIOR BELIEF** in quanto sono probabili i vari valori che \bar{w} può assumere.

Per rendere la priori, possiamo assumere che i pesi \bar{w} assumano MEDIANTE UNO INTRICO ALCI ZERO, con una CERTA VARIANZA ALCI' ESSA INTRO ALLO ZERO.

$$P(\bar{w}|\alpha) = N(\bar{w}|0, \alpha^{-1} I)$$

E' la priori assunta INIZIALE

possiamo discutere la priori verso le forme di una
GAUSSIAN CONJUGATE PRIOR, il che significa che possiamo
 utilizzare la stessa priori per la **LIAELHOOD**
GAUSSIANA la distribuzione a posteriori risulta essere
 anche essa una **GAUSSIANA!**

$$p(\bar{w} | \bar{t}) = p(\bar{t} | \bar{w}, \beta^{-1}) p(\bar{w} | \alpha) \sim N(\bar{w} | m_n, s_n)$$

DENSITÀ
LIAELHOOD
→ PRIOR

DISTRIBUZIONI
A POSTERIORI

Dove i valori delle norme sono rispettivamente:

- $m_n = \beta S_n \phi^T \bar{t}$ / media
- $S_n^{-1} = \alpha I + \beta \phi \phi^T$ / varianza

possiamo semplificare ancora a calcolare il \ln :

$$\ln p(\bar{w} | \bar{t}) = -\frac{\beta}{2} \sum_{i=1}^N \left\{ t_i - \underbrace{\bar{w}^T \phi(\bar{x}_i)} \right\}^2 - \frac{\alpha}{2} \bar{w}^T \bar{w} + costante$$

→ model prediction

possiamo quindi notare che risulta la **DISTRIBUZIONE A POSTERIORI MASSIMA (MAP)**, ovvero $\max \ln p(\bar{w} | \bar{t})$ oppure a
 equivalente $\min \frac{1}{2} \bar{w}^T \bar{w} - \sum_{i=1}^N \bar{w}^T \phi(\bar{x}_i) + \text{costante}$ con valore visto nel

REGRESSONE LINEARE STANDARIZZATA CON FATTORE DI REGOLAZIONE:

$$\lambda = \frac{\alpha}{\beta}$$

STANZIAMO PUNTI, NEL CASO PUÒESSERE DI PIÙ POTENTI,
AVETE POSSIAMO ESCREIRE UNA REGRESSIONE PUNTIPLICANDO IL
BESTE IN CUI SOLUZIONE DEI PARAMETRI W* RISULTA!

INOLTRE, POSSIAMO PERMETTERE DI "IMPOSTARE IN MODO
INCREMENTALE", OVETRO NELL'ESO IN CUI SIAMO A DISPOSIZIONE
NUOVE INFORMAZIONI SUI DATI, BASTERA' AGGIORNARE AI DATI
CHE SONO PRESENTI INCORPORANDO IN NUOVA CONOSCENZA.

- > SI CREA UNA **LINEA** SUI DATI PRESENTI
- > SI IMPOSTA UNA **PROB. INIZIALE** CON NUOVI
E VARIABI FISSATI
- > SI MOLTIPLICA **LINCUSSIONE X PROB. OTTENUTA**
CON **POSTERIORI CHE FORNISCE LA DISTRIBUZIONE**
DEI NUOVI DATI NEGLI MODELLI GDI SI POSSO BREVIP
- > NELL'ESO DI AGGIUNTA DI NUOVI DATI IN INPUT
SI AGGIORNÀ LA PROB. CON LE NUOVE INFORMAZIONI
OTTENENDO QUINDI UNA NUOVA POSTERIORI E COSÌ VAI

TRAMITE QUESTO PROCESSO CI VIVEREMO, CON INIZIALMENTE

È un modo di pensare perché si mette molto informazioni, se usi un valore e ci si basa su uno su cui **stiamo** nella **posterior** che si avanza al valore effettivo della **maximum likelihood**, con in più l'informazione sul **bayes**.

Per quanto riguarda le **precisioni** del modello, se tu sei preoccupato per l'**output** $y(\bar{e}, \bar{w})$ quanto è **corretto**? In pratica non ti interessa il valore effettivo di \bar{w} , ma vorrai solo ottenere delle **distribuzioni** di **precisioni**:

$$p(t | \bar{e}, \alpha, \beta) = \int p(t | \bar{w}, \beta) p(\bar{w} | \bar{e}, \alpha, \beta) d\bar{w}$$

↳ **value?**

↳ **likelihood** ↳ **posterior**

Questa formula rappresenta la **MEDIA DELLE LIKELIHOOD CONDIZIONATE** rispetto al valore della **posterior**, dove una distribuzione dei parametri. Inoltre, essendo che entrambi sono delle distribuzioni normali, si ha che la distribuzione della predizione è la **convoluzione di due normali**:

$$p(t | \bar{e}, \alpha, \beta) = N(t | m_n \phi(\bar{x}), \sigma_n^2(\bar{x}))$$

Dove $\sigma_n^2(\bar{x}) = \frac{1}{\beta} + \phi(\bar{x})^\top S_n \phi(\bar{x})$ e dove $\frac{1}{\beta}$ rappresenta

il numero delle **informazioni**, mentre σ_n^2 rappresenta il **livello di incertezza** nelle stime dei parametri.

