2/11/15

GAUSSIANS & REGRESSION

① High-level Picture & Flow of topic in Class

	Last Time	Today / Next lecture
Real-World Problem	Binary / Discrete decisions: Nadal W/L, Email Spam/Non-spam	Continuous-valued variables: House prices, stock index, etc
Variable :	$Y = y \in \{0,1\}$ [or $\{1,\dots k\}$]	$Y = y \in \mathbb{R}$ [later $\mathbb{R}^d$]
Model :	$Y \sim Bea(\theta)$ [or $Y \sim Cat(\vec{\theta})$]	$Y \sim N(\mu, \sigma^2)$
Parameters :	$\theta = P(Y=1) = E[\mathbb{I}[Y=1]]$	$\mu = E[Y]$ $\quad \sigma^2 = Var(Y) = E[(Y-\mu)^2]$
MLE :	$\hat{\theta}_{MLE} = \frac{\#(H)}{\#H + \#T}$	$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} y_i$ $\quad \hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})^2$
Prior :	$\theta \sim Beta(\beta_H, \beta_T) \Leftrightarrow P(\theta \mid \beta_H, \beta_T) \propto \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}$	$\mu \sim N(\eta, \lambda^2)$ $\quad \sigma \sim$ Inverse Gamma

These are all models for Y (not too interesting)
We will later move to models of $Y \mid X = x$
(Y: output, x: input / features)

	$Y \mid X = x \sim Bea(\theta_x)$ (depends on input)	$Y \mid X = x \sim N(\mu_x, \sigma_x^2)$
	classification	regression
	{Logistic Regression, Naive Bayes}	{Least Squares Regression}

(2) Meet the Gaussian Density Function

$$Y \sim N(\mu, \sigma^2)$$

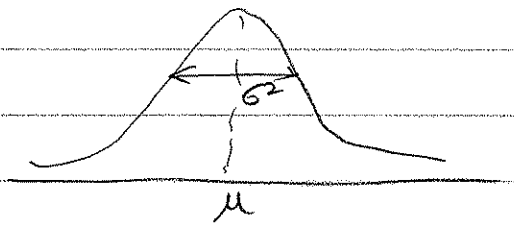
$$\equiv p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

Recall: Density function: $\int_{-\infty}^{\infty} p(y)\,dy = 1$

$$Pr[a \le Y \le b] = \int_a^b p(y)\,dy$$

$$E[f(Y)] = \int_{-\infty}^{\infty} f(y)\,p(y)\,dy$$

$p(y) \ge 0$ but NOT necessarily < 1



$\mu = E[Y]$

$\sigma^2 = E[(Y-\mu)^2]$

(3) Max-likelihood Estimation of μ, σ (Same Principles)

Dataset $D = \{y_1, \ldots, y_N\}$ (may be exam scores of students) $\quad Y \sim N(\mu, \sigma^2)$

$$(\hat{\mu}_{MLE}, \hat{\sigma}_{MLE}) = \underset{\mu,\sigma}{\operatorname{argmax}}\ LL(\mu, \sigma)$$

[Cranking through the math]

$$LL(\mu,\sigma) = \log P(D \mid \mu,\sigma)$$

$$= \log \prod_{i=1}^{N} p(y_i \mid \mu,\sigma) \qquad [\text{IID}]$$

$$= \sum_{i=1}^{N} \log p(y_i \mid \mu,\sigma)$$

$$= \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\mu)^2}{2\sigma^2}} \right]$$

$$= \sum_{i=1}^{N} -\frac{(y_i-\mu)^2}{2\sigma^2} - \frac{N}{2}\log(2\pi\sigma^2)$$

Now, Calculus!

$$\frac{\partial LL}{\partial \mu} = -\sum_{i=1}^{N} \frac{2(y_i-\mu)(-1)}{2\sigma^2} = 0$$

$$\Rightarrow \sum_{i=1}^{N} (y_i-\mu) = 0 \Rightarrow \boxed{\hat{\mu}_{MLE} = \frac{1}{N}\sum y_i}$$

$$\frac{\partial LL}{\partial \sigma} = -\sum \frac{(y_i-\mu)^2}{2}\left(\frac{-2}{\sigma^3}\right) - \frac{N}{2}\frac{1}{2\pi\sigma^2}\, 2\pi\cdot(2\sigma) = 0$$

$$= \sum_{i=1}^{N} \frac{(y_i-\mu)^2}{\sigma^3} - \frac{N}{\sigma} = 0$$

$$\Rightarrow \sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{\mu}_{MLE})^2 \qquad \{\text{assuming } \sigma \neq 0\}$$

(4) MAP estimation of μ

We said $\{y_1, \ldots, y_N\}$ may be exam scores of students (say)

$P(\mu)$ = I know I set the exam hard
Class average should be around 20/100
(Just a thought. Don't panic)

For simplicity assume σ = (unknown) constant

$$\mu \sim N(\eta, \lambda^2)$$

η, λ: hyper-parameters [known!]

Why another Gaussian?
It's self-conjugate!
(aka math is simple)

$$\hat{\mu}_{MAP} = \underset{\mu}{\operatorname{argmax}} \log P(\mu | D)$$

$$= \underset{\mu}{\operatorname{argmax}} \log \frac{P(D|\mu) P(\mu|\eta,\lambda^2)}{P(D)}$$

$$= \underset{\mu}{\operatorname{argmax}} \log P(D|\mu) + \log P(\mu|\eta,\lambda^2) - \underbrace{\log P(D)}_{\text{const w.r.t } \mu}$$

$$\frac{\partial [\log P(\mu|D)]}{\partial \mu} = \frac{\partial LL}{\partial \mu} + \frac{\partial}{\partial \mu} \log p(\mu|\eta,\lambda^2)$$

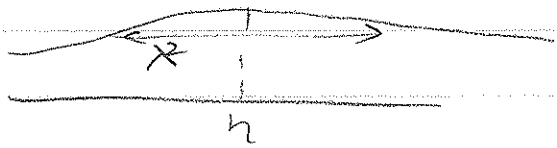
$$= \sum_{i=1}^{N} \frac{(y_i - \mu)}{\sigma^2} + \frac{\partial}{\partial \mu}\left[-\frac{(\mu-\eta)^2}{2\lambda^2} - \frac{1}{2}\log(2\pi\lambda^2)\right]$$

$$= \sum_{i=1}^{N} \frac{(y_i - \mu)}{\sigma^2} - \frac{(\mu-\eta)}{\lambda^2} = 0$$

$$\Rightarrow \boxed{\hat{\mu}_{MAP} = \frac{\left(\frac{\sum_{i=1}^{N} y_i}{\sigma^2} + \frac{\eta}{\lambda^2}\right)}{\frac{N}{\sigma^2} + \frac{1}{\lambda^2}}}$$

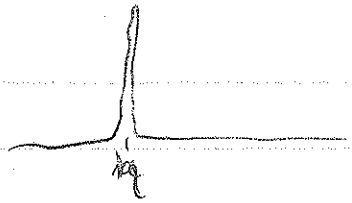
→ Consider, $\lambda \to \infty$ $\quad p(\mu|\eta,\lambda^2) = \frac{1}{\sqrt{2\pi}\lambda^2} e^{-\frac{(\mu-\eta)^2}{\lambda^2}}$

$p(\mu|\eta,\lambda)$ is nearly unif



$$\hat{\mu}_{MAP} \to \frac{\sum_{i=1}^{N} \frac{y_i}{\sigma^2} + (\to 0)}{\frac{N}{\sigma^2} + (\to 0)} \to \hat{\mu}_{MLE} \quad \text{(nice)}$$

→ Consider $\lambda \to 0$ $\quad p(\mu|\eta,\lambda^2)$



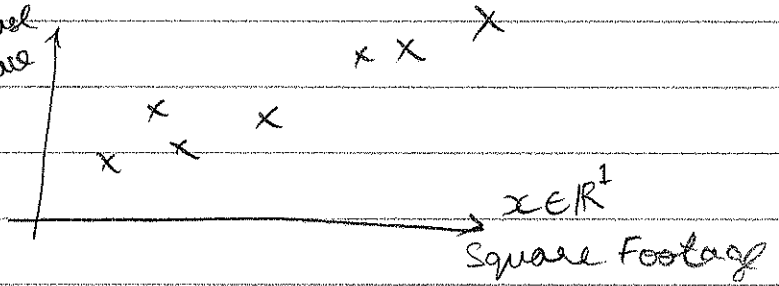
$$\hat{\mu}_{MAP} \cong \frac{\lambda^2 \sum \frac{y_i}{N} + \eta}{\lambda^2 \cdot \frac{N}{\sigma^2} + 1}$$

→ $\hat{\mu}_{MAP} \to \eta$ (prior mode)

(5) Regression

Task: Prediction continuous values $X \to Y \in \mathbb{R}$

Example



Dataset: $D = \{(\vec{x}_1, y_1), \ldots (\vec{x}_n, y_n)\}$

$\vec{x} \in \mathbb{R}^d$ (Input / features)

$y \in \mathbb{R}$ (Output)

Model: $\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d$

$$= [w_0 \ldots w_d] \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$= w^T \vec{x}$

or $\vec{x}^T w$

(6) Least-Square Fitting

Error / Residuals $e_i = y_i - \hat{y}_i$

y_i: Ground Truth; $\hat{y}_i$: prediction

Natural Loss Function $L(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$

Squared Loss

So our objective function / Loss function on the dataset

$$\min_{\vec{w} \in \mathbb{R}^{d+1}} L(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \vec{w}^T\vec{x}_i)^2$$

Now we could view $L(\vec{w}) \equiv L(w_0, w_1, \ldots, w_d)$ to start chugging along standard calculus

$$\left.\begin{aligned} \frac{\partial L}{\partial w_0} &= 0 \\ \frac{\partial L}{\partial w_1} &= 0 \\ &\vdots \\ \frac{\partial L}{\partial w_d} &= 0 \end{aligned}\right\} \text{ solve system of equations}$$

But let's make the notation more compact with matrices

$X_{n\times(d+1)}$ = data matrix / "design matrix"

$$= \begin{bmatrix} \leftarrow \vec{x}_1^T \rightarrow \\ \vdots \\ \leftarrow \vec{x}_n^T \rightarrow \end{bmatrix}_{n\times(d+1)}$$

each training instance is a row

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$ all GTs / labels

$$\hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}$$ all prediction

$$= \begin{bmatrix} \longleftarrow \vec{x}_i \longrightarrow \end{bmatrix} \times \begin{bmatrix} w_0 \\ \vdots \\ w_d \end{bmatrix} \quad \vec{w} = Xw$$

So $$L(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \frac{1}{n} \| Y - \hat{Y} \|_2^2$$

$$= \frac{1}{n} (Y - \hat{Y})^T (Y - \hat{Y})$$

$$= \frac{1}{n} (Y - Xw)^T (Y - Xw)$$

$$= \frac{1}{n} \left[Y^T Y - 2 Y^T Xw + w^T X^T Xw \right]$$