Beta Prior Distribution

$$P(\theta) = \frac{\theta^{B_H - 1}(1-\theta)^{B_T - 1}}{B(B_H, B_T)}$$

covers only material covered through today

Conjugate Priors

$P(\theta|D)$ is in the same distribution as $P(\theta)$.

ex $P(\theta) \sim$ Beta, $P(\theta|D) \sim$ Beta

Conjugate Prior $P(\theta)$

specify which distribution it follows

Ex 1

is mean and mode same

mean $\frac{B_H}{B_H + B_T}$ Mode $\frac{B_H - 1}{B_H + B_T - 2}$

mean = mode iff $B_H = B_T$

$$P(D|\theta) \cdot P(\theta) = \overbrace{\theta^{B_H - 1} \cdot (1-\theta)^{B_T - 1} \cdot \frac{1}{B(B_H, B_T)}}^{\text{Prior}} \cdot \underbrace{\theta^{\alpha_H}(1-\theta)^{\alpha_T}}_{\text{likelihood}}$$

$$= \frac{1}{B(\cdot)} \cdot \theta^{\alpha_H + B_H - 1}(1-\theta)^{\alpha_T + B_T - 1}$$

Posterior = $P(\theta|D) = \frac{1}{C} P(D|\theta) P(\theta)$

$$= \frac{1}{C} \frac{1}{B(\cdot)} \theta^{\alpha_H + B_H - 1}(1-\theta)^{\alpha_T + B_T - 1}$$

$\sim$ Beta$(\alpha_H + B_H, \alpha_T + B_T)$

// How does the $\frac{1}{C}$ factor in?

Map = mode

$\theta | D \sim$ mean $\quad \frac{\alpha_H + \beta_H}{\alpha_H + \beta_H + \alpha_T + \beta_T}$

MLE $\quad \frac{\alpha_H}{\alpha_H + \alpha_T}$

Ex 2 MAP for Beta

$p(\theta | D)$

// How is MLE different from MAP

Prior: $\theta \sim Beta(2,3)$

① 2 heads 2 tails

$\hat{\theta}_{MLE} = \frac{2}{2+2} = \frac{1}{2}$

$\hat{\theta}_{MAP} = \frac{4-1}{4+5-2} = \frac{3}{7}$

$p(\theta | D) \sim B(4,5) = B(2+2, 3+2)$

② 200 heads 200 tails

$\hat{\theta}_{MLE} = \frac{200}{200+200} = \frac{1}{5}$

$\hat{\theta}_{MAP} = \frac{200-1+2}{202+203-2} = \frac{201}{403}$

Ex

posterior $B(30,40)$ Prior $B(2,3)$

28 H 37 T $\quad B_H = 2 \quad B_T = 3$

Posterior: $B(B_H+\alpha_H, B_T+\alpha_T)$

$30 = 2+\alpha_H, \; 3+\alpha_T = 40)$

$28 = \alpha_H \quad \alpha_T = 37$

Gaussians Properties Pg 19

How do we learn μ, σ

MLE for Gaussian

$$P(D|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\ln(P(D|\mu,\sigma)) = \ln \; ''$$

$$= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

MLE for μ (mean) hold σ constant

$$\frac{d}{d\mu}\ell(\mu,\sigma) = -\cancel{N \ln\sigma\sqrt{2\pi}}^{0} - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= + \sum_i 2\,\omega =$$

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

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

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

$$= \sum_{i=1}^{n} \frac{x_i-\mu}{\sigma^2} \cdot \sigma^2 = 0 \cdot \sigma^2$$

$$= \sum_{i=1}^{n} x_i - \mu = 0$$

$$= \sum_i (x_i) - N\mu = 0 \quad // \mu \text{ is constant}$$

$$\sum_i (x_i - \mu) = 0$$

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_i x_i$$

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_i x_i$$

$E(\hat{\mu}_{MLE})$

$$E(\hat{\mu}_{MLE}) = \frac{1}{N}\sum_i E(x_i) \qquad \boxed{\frac{1}{N}\sum_i^N (x_i) = \mu}$$

$$= \frac{1}{N}\sum_i \mu$$

$$= \mu \quad \text{// true parameter of mean}$$

$$\Rightarrow \hat{\mu}_{MLE} \text{ is unbiased}$$

Properties of MLE Fisher info
↑ OPTIONAL ↑

MLE Variance p24

$$\frac{\partial}{\partial \sigma}\left(-N \ln \sigma\sqrt{2\pi} - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= +\frac{N}{\sigma} + \sum_i \frac{(x_i-\mu)^2}{\sigma^3}$$

$$-N \ln \sigma\sqrt{2\pi} - 2^{-1}\sigma^{-2}\sum_i (x_i-\mu)^2$$

$$= \frac{-N}{\sigma\sqrt{2\pi}}\cdot\sqrt{2\pi} + 2^{-1}(2)\sigma^{-3}\sum_i (x_i-\mu)^2$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_i (x_i-\mu)^2 \qquad 0 = \frac{-N}{\sigma} + \sum_i \frac{(x_i-\mu)^2}{\sigma^3} \Rightarrow \frac{N\cdot\sigma^3}{\sigma} = \sum_i (x_i-\mu)^2$$

$$\Rightarrow \hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i-\mu)^2$$

⇒ use MLE for μ since we do not have true ⇒ sample variance

$\hat{\sigma}^2_{MLE}$ is biased. Why? It uses an estimated $\hat{\mu}$

$$\hat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum(x_i-\hat{\mu})^2$$

$$= \frac{1}{N}\sum(x_i-\mu)^2$$

check $E(\hat{\sigma}^2_{MLE})$
should be σ^2
if not, then biased

Bayesian Learning of Gaussian Distribution

Conjugate Prior

Mean Gaussian Prior

Variance Wishart Distribution ← Next covered

Prior for mean

$$p(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda}}$$

η, λ hyperparameters

MAP Gaussian Mean

$$p(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(\mu-\eta)^2}{2\lambda^2}}$$

likelihood * prior

$$p(D \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$\frac{d}{d\mu}[(\ln p(D \mid \mu, \sigma^2) \cdot p(\mu \mid \eta, \lambda)]$ argmax$_\mu$ MLE → MAP for μ

$$\ln \frac{1}{\lambda\sqrt{2\pi}} - \frac{(\mu-\eta)^2}{2\lambda^2} + N \ln \frac{1}{\sigma\sqrt{2\pi}} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

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

$$-\frac{1}{\lambda^2}\mu + \frac{\eta}{\lambda^2} + \frac{1}{\sigma^2}\sum_i x_i - \frac{N}{\sigma^2}\mu = 0$$

$$\hat{\mu}_{MAP} = \frac{\left(\frac{1}{\sigma^2}\sum x_i + \frac{\eta}{\lambda^2}\right)}{\left(\frac{N}{\sigma^2} + \frac{1}{\lambda^2}\right)}$$

$\frac{1}{N} \sum_i x_i$ = mean

Frequentist slides
Bayesian Statistics
probability as degrees of belief

4:21 pm of the last day:
we start on Regression

Regression Minimize
$w^* = E$: residual
$= \text{argmin}_w (Hw-t)^T(Hw-t)$
$\frac{d}{dw} = w^T H^T H w - 2t^T H w + t^T t$

$2H^T Hw - 2H^T t = 0$
$H^T Hw = H^T t$ Pseudo inverse ?
$w^* = (H^T H)^{-1} H^T t$

Maximize log-likelihood

$$\ln(P(D|w,\sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(t_i - \sum w_i h_i(x_i))^2}{2\sigma^2}}$$

assumed noise $\sim N(0,1)$

$$\frac{\partial}{\partial w} \ln\left(\prod_{i=1}^{N} e^{-\frac{(t_i - \sum w_i h_i(x_i))^2}{2\sigma^2}}\right)$$

$$= \frac{2}{2w} \sum_{i=1}^{N} -\frac{(t_i - \sum w_i h_i(x_i))^2}{2\sigma^2}$$

$$= \sum (t_i - \sum w_i h_i(x_i))^2 \leftarrow \begin{matrix}\text{residual} \\ \text{or} \\ \text{SSE}\end{matrix}$$

Least-squares Linear Regression for Gaussians

linear algebra slides for offline review

link for regression on BB next week

SVD not on First midterm

- skip lecture 3 -

Start Lecture 4

Theorem: Bayes classifier is optimal

h_{Bayes}

$error_{True}(h_{Bayes}) \leq error_{True}(h), \forall h(x)$ ← this is provable

proof: $p(error) = \int_x p(error(x) p(x)) dx$

Bayes classifier

$$h(x) = \begin{cases} 1 & \text{if } p(y=1) \geq 0.5 \\ -1 & \text{if } p(y=1) < 0.5 \end{cases}$$

Bayes Rule

$$p(y|x) = \frac{p(x|y)\, p(y)}{p(x)}$$

likelihood, prior, marginal

How hard to learn optimal classifier

$x \in R^6$

Prior $P(y)$ of K classes (y_1, y_0)

if $K = 2$, 1 param needed

else need $K-1$ parameters

n features in m-space

$p(x|y)$ where x composed of n binary features

DISCRETE: $2^n \cdot ($ how does binary take effect?

$m^n - 1$ possible features

$2^1 - 1 = 1$ due to binary choice

$p(x=1 \mid y=1) = 1 - \text{"} = p(x=0)$

$(m^n - 1)^{k-1}$

for k classes
and n binary ($m=2$) features

prior $p(y)$ $(k-1)$ parameters
for each y, $p(x \mid y=k)$
$(m^n - 1)$ parameters
$k(m^n - 1) + k - 1$ parameters over all
exponential in terms of the number
of features.

Naive Bayes Assumption

Assume features are conditionally independent
$p(E, A \mid P_a) = p(E \mid P_a) \cdot p(A \mid P_a)$
features class

new feature space of $p(x \mid y=1)$ // no effect on $p(y)$
n features (binary) no correlation

Kn factors $\forall y$

$$y^* = h_{NB}(x) = \arg\max_y P(y) P(x_{1:N} | y)$$
$$= \arg\max_y P(y) \prod_i p(x_i | y)$$

via conditional independence assumption

if assumption holds, NB optimal
usually not true, but good enough

MLE for NB

Prior $P(Y=y)$

$$\frac{y}{\sum_i y_i}$$

Likelihood $P(X_i = x_i | Y_j = y_j)$

$$P(X_i = x_i | Y_i = y_i) = \frac{\#X_i = x_i \text{ and } Y_i = y_i}{\# Y_i = y_i}$$

No MAP on Exam