Today - Beta - Bernoulli Model (for binary data)

- Dirichlet Model (for multiclass data)

- **Naive Bayes classifier**

Problem - email classification → { regular, important, spam } classes

- Input to Naive Bayes - 1) prior over classes

2) dataset of labelled examples

$(D_1, c_1), (D_2, c_2) \ldots (D_n, c_n)$

D_i ↑ email ≡ 'bag of words'

3) new email D

⇒ output ≡ distn over $C(D)$

'Hi
Lets meet after class'

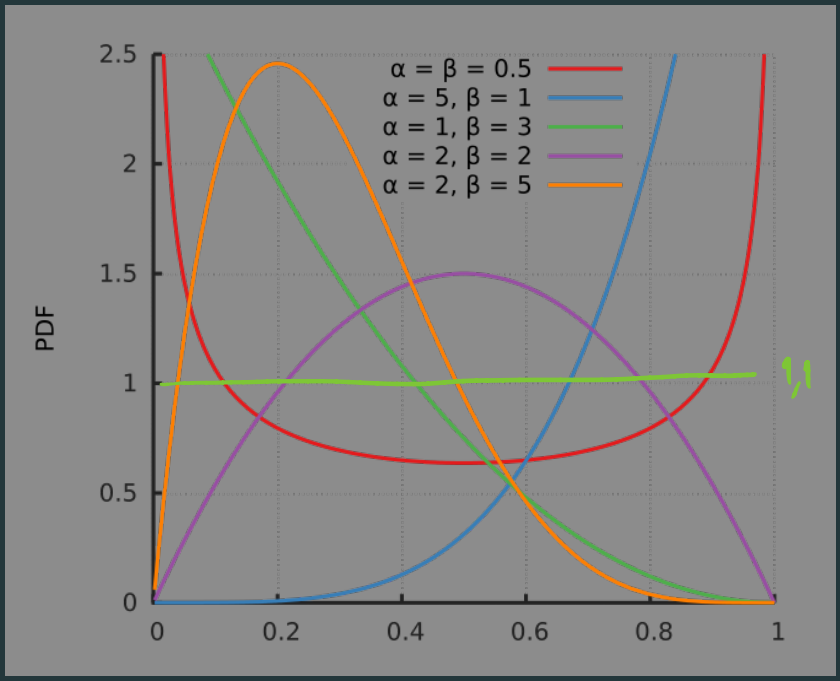
⇒

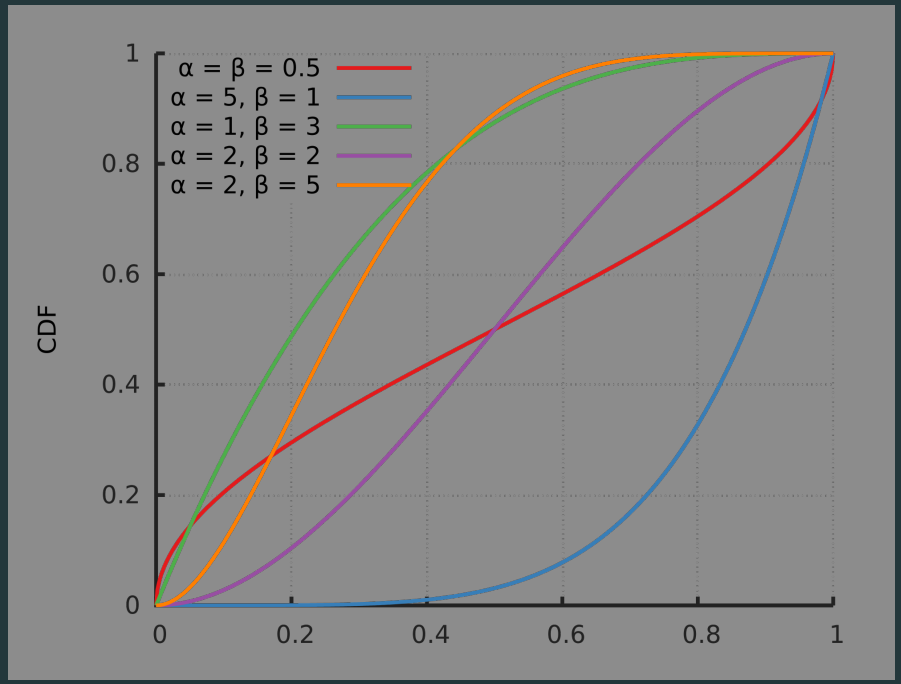
Hi	1
meet	1
lets	1
.	
.	
.	

the Beta distribution

Beta distribution

- $x \in [0, 1]$, parameters: $\Theta = (\alpha, \beta) \in \mathbb{R}^+$ ('# ones'+1,'# zeros'+1)
- pdf: $p(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$ ⟵ same form as the Bernoulli likelihood!
- normalizing constant: $\frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} = \frac{1}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx}$ $= \frac{1}{B(\alpha,\beta)}$





Beta-Bernoulli prior and updates

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

Beta-Bernoulli model

- prior parameters: $\Theta_0 = (\alpha, \beta) \in \mathbb{R}^+$ (hyperparameters)
- Beta-Bernoulli prior: $Beta(\alpha, \beta) \sim p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- likelihood: $p(D|\theta) = \theta^{N_1}(1-\theta)^{N_0}$

  then via Bayesian update we get $p(\theta|D) = L(\theta|D)\, p(\theta)$

- posterior:

$$p(\theta|D) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{N_1}(1-\theta)^{N_0} \sim Beta(\alpha + N_1, \beta + N_0)$$

the Beta distribution: getting familiar

$Beta(\alpha, \beta)$ distribution

$$p(x) = \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}_{\frac{1}{B(\alpha,\beta)}} x^{\alpha-1}(1-x)^{\beta-1}, \quad B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

properties of $\Gamma(\alpha)$

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1}\,dy$$

more importantly

$$\boxed{\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)} \qquad \boxed{\Gamma(1)=1}$$

- If α is an integer $\Rightarrow \Gamma(\alpha) = (\alpha-1)!$

the Beta distribution: mean and mode

$Beta(\alpha, \beta)$ distribution

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

- $\mathbb{E}[X] = \int_0^1 x \cdot \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx$

$= \frac{1}{B(\alpha,\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}\,dx = \frac{B(\alpha+1,\beta)}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)}$

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

$\Rightarrow$ $\boxed{Mean\left(Beta(\alpha,\beta)\right) = \frac{\alpha}{\alpha+\beta}}$

$$\text{mode} \equiv \underset{x \in [0,1]}{\arg\max} \; \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$$

$$\frac{d}{dx}\left(x^{\alpha-1}(1-x)^{\beta-1}\right) = (\alpha-1)x^{\alpha-2}(1-x)^{\beta-1} - (\beta-1)x^{\alpha-1}(1-x)^{\beta-2}$$

$$= x^{\alpha-2}(1-x)^{\beta-2}\Big((\alpha-1)x - (\beta-1)(1-x) = 0$$

$$\Rightarrow \quad x^* = 0, 1 \text{ or } \frac{\alpha-1}{\alpha+\beta-2}$$

Mode of Beta(α,β) dist^n is $\dfrac{\alpha-1}{\alpha+\beta-2}$

Beta-Bernoulli model: what should we report?

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

prior hyperparams

posterior hyperparams

Correct answer $\equiv$ report (prior + posterior) $\equiv Beta(\alpha+N_1, \beta+N_0)$

- In practice – 'people' want a single number

For this – decision theory –

- ask what the 'loss function' is
- return the θ which minimizes expected loss over the posterior

decision theory

Idea - Choose 'action' to minimize expected loss over $p(\theta|D)$

Eg - Let θ be the 'true answer', $\hat{\theta}$ be the report

1) $L(\theta, \hat{\theta}) = \mathbb{1}_{\{\theta \neq \hat{\theta}\}}$ (L_0 loss) : $\hat{\theta}$ = mode of $p(\theta|D)$

2) $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ (L_2 loss) : $\hat{\theta}$ ≡ mean of $p(\theta|D)$

3) $L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$ (L_1 loss) : $\hat{\theta}$ ≡ median of $p(\theta|D)$

general rule - $\hat{\theta} = \operatorname{argmin} \mathbb{E}_{p(\theta|D)}[L(\theta, \hat{\theta})]$

decision theory in a nutshell

Bayesian decision theory in learning

given prior F on θ, choose 'action' $\hat{\theta}$ to minimize loss function $\mathbb{E}_F[L(\theta, \hat{\theta})]$

examples

- L_0 loss: $L(\theta, \hat{\theta}) = \mathbb{1}_{\{\theta \neq \hat{\theta}} \Rightarrow \hat{\theta_{L_0}} =$ mode of F
- L_1 loss: $L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_1 \Rightarrow \hat{\theta_{L_1}} =$ median of θ under F
- L_2 loss: $L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_2 \Rightarrow \hat{\theta_{L_2}} = \mathbb{E}_F[\theta]$

in general 'decision-making'

given prior F on X, choose 'action' $a \in \mathcal{A}$ to minimize loss, i.e.

$$a^* = \arg\min_{a \in \mathcal{A}} \mathbb{E}_{X \sim F}[L(a, X)]$$

Beta-Bernoulli model: posterior mean

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

posterior mean: $\mathbb{E}[\theta|\alpha, \beta, N_0, N_1] = \mathbb{E}[Beta(\alpha+N_1, \beta+N_0)] = \dfrac{\alpha+N_1}{\alpha+\beta+N_0+N_1}$

Defn - $n = N_1 + N_0$

$m = \alpha + \beta$

m = '# of prior samples'

$\frac{\alpha}{m}$ = prior mean

$\frac{N_1}{n}$ = Data mean (also, MLE)

$w = \frac{m}{m+n}$ 'strength of prior'

$$= \frac{\alpha+N_1}{m+n} = \left(\frac{\alpha}{m}\right)\cdot\left(\frac{m}{m+n}\right) + \left(\frac{N_1}{n}\right)\left(\frac{n}{m+n}\right)$$

$$= \underbrace{(\text{prior mean})\cdot w}_{\text{regularizer}} + (\text{MLE})\underbrace{(1-w)}_{\text{shrinkage}}$$

Note: For fixed m, as $n \to \infty$, then Posterior mean $\longrightarrow$ MLE

Beta-Bernoulli model: posterior mode (MAP estimation)

max a posteriori

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

posterior mode: $\max_{\theta \in [0,1]} p(\theta|\alpha, \beta, N_0, N_1) = \frac{\alpha + N_1 - 1}{\alpha + \beta + N_1 + N_0 - 2} = \frac{\alpha + N_1 - 1}{n + n - 2}$

- If $\alpha = \beta = 1$ (uniform prior)

$$\Rightarrow \quad \theta_{MAP} = \frac{N_1}{n} = \theta_{MLE}$$

i.e.: MLE is the 'right' answer if L0 loss + uniform prior

Beta-Bernoulli model: posterior prediction (marginalization)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution
- prior: $p(\theta) \sim Beta(\alpha, \beta)$ posterior: $p(\theta|D) \sim Beta(\alpha + N_1, \beta + N_0)$

posterior prediction: $\mathbb{P}[X = 1|D] = \int_0^1 p(\theta|D) \cdot \theta \cdot d\theta$

$$= \frac{\alpha + N_1}{n + n}$$

- If $\alpha = \beta = 1$

$$\mathbb{P}[X=1|D] = \frac{N_1 + 1}{N_0 + N_1 + 2}$$

Laplace Estimator

If $N_1 = 0$, $\mathbb{P}[X=1|D] = \frac{1}{n+2}$

the black swan

- In some datasets, a rare event is never seen
  - If n samples, $N_1 = 0$, $N_0 = n$
    then Laplace estimator $\mathbb{P}[X=1 \mid \mathcal{D}] = \frac{1}{n+2}$ (prob for the 'unseen')

$$\mathbb{P}[X=0 \mid \mathcal{D}] = \frac{n+1}{n+2}$$

- One important application $\equiv$ crypto
  (Good-Turing estimator)

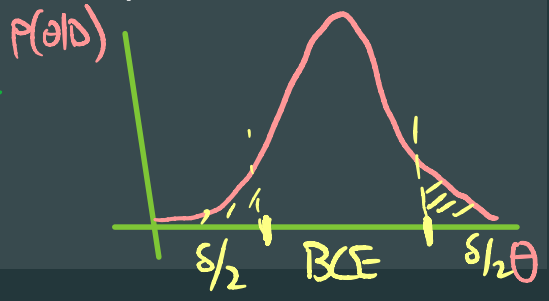
summarizing the posterior

model $\mathcal{M}+$ prior $p(\Theta)+$ data $D \Rightarrow$ posterior $p(\Theta|D)$

summarizing $p(\Theta|D)$

- posterior mean $\widehat{\theta}_{mean} = \mathbb{E}[\Theta|D]$
- posterior mode (or MAP estimate) $\widehat{\theta}_{MAP} = \arg\max_{\Theta} p(\Theta|D)$
- posterior median $\widehat{\theta}_{median} = \min\{\Theta : p(\Theta|D) \geq 0.5\}$
- Bayesian credible intervals: given $\delta > 0$, want (ℓ_Θ, u_Θ) s.t.

$$\mathbb{P}[\ell_\Theta \leq \Theta \leq u_\Theta|D] > 1 - \delta$$



marginal likelihood (evidence)

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{0,1\}^n$, contains N_1 ones and N_0 zeros
- model $\mathcal{M}$: X_i are generated i.i.d. from a $Ber(\theta)$ distribution

marginal likelihood

$$p(D) = \frac{p(\theta)p(D|\theta)}{p(\theta|D)} = \frac{\text{prior} \times \text{likelihood}}{\text{posterior}}$$

$$P(D) = \frac{\Gamma(m)}{\Gamma(n+m)} \cdot \frac{\Gamma(N_1+\alpha)}{\Gamma(\alpha)} \cdot \frac{\Gamma(N_0+\beta)}{\Gamma(\beta)}$$

- Q: Given data, is it from $Ber(0.5)$ vs $Ber(p)$ for $p \notin [0.5-\delta, 0.5+\delta]$

'marginal likelihood lets you compare models'

multiclass data

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{1, 2, \ldots, K\}^n$
- for $i \in [K]$, data D contains N_i copies of type i
- model $\mathcal{M}$: X_i generated i.i.d. from $\mathit{Multinomial}(\theta_1, \theta_2, \ldots, \theta_K)$ distn