What is Dirichlet dist.?

→ Start with Beta dist:

flip a coin w. prob. q turning up head

prob. of x heads in n flips → Binomial dist.

$$p(x|q,n) = \binom{n}{x} q^x (1-q)^{n-x}$$

what if you don't know q?

→ Bayesians represent uncertainty about q with a prior dist. (i.e. Beta dist.)

$$p(q|\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} q^{\alpha_1 - 1} (1-q)^{\alpha_2 - 1}$$

hyperparameters $\alpha_1, \alpha_2 > 0$ ← are pseudocoun

Why Beta?

- it is defined over interval [0, 1
- Beta and Binomial are conjugat (a Beta prior and Binomial likeli result in a posterior Beta)

posterior $\propto$ likelihood $\times$ prior

$$p(q|x,\alpha_1,\alpha_2) \propto p(x|q,n)\, p(q|\alpha_1,\alpha_2)$$

$$\propto q^x (1-q)^{n-x} q^{\alpha_1 - 1} (1-q)^{\alpha_2 - 1}$$

$$= q^{\alpha_1 + x - 1} (1-q)^{\alpha_2 + n - x - 1}$$

$$= \text{Beta}(\alpha_1 + x, \alpha_2 + n - x)$$

DP-03

Dirichlet (Multinomial conjugacy)

(٢)

- Multinomial extends binomial to more the 2 classes
(instead of flipping a coin, roll a die)

Let Z be the multinomial random var. with $p(Z_K=1)=q_K$

- Dirichlet dis. is conjugate prior to multinomial

- It is defined as an exponential family dist. on the simplex (generalizes the Beta)

Simplex:
generalization of a triangle to arbitrary dimension
K-simplex is a K-dim. polytope which is convex hull of its K-1 verticies.

$$p(q|\alpha)=\frac{\Gamma\left(\sum_{i=1}^{m}\alpha_i\right)}{\prod_{i=1}^{m}\Gamma(\alpha_i)}q_1^{\alpha_1-1}\cdots q_m^{\alpha_m-1}$$

where $q=(q_1,q_2,\ldots q_m)$ is a point on $(m-1)$-simplex

(i.e. $0<q_i<1$ and $\sum_{i=1}^{m}q_i=1$

and $\alpha=(\alpha_1,\ldots\alpha_m)$ is a set of parameters $(\alpha_i>0)$

posterior $\propto$ likelihood $\times$ prior

$$p(q|z,\alpha)\propto p(z|q)\,p(q|\alpha)$$

$$\propto(q_1^{z_1},\ldots,q_m^{z_m})(q_1^{\alpha_1-1},\ldots q_m^{\alpha_m-1})$$

$$\propto(q_1^{\alpha_1+z_1-1},\ldots,q_m^{\alpha_m+z_m-1}$$

$$=Dir(\alpha_1+z_1,\ldots\alpha_m+z_m)$$

Properties of Dirichlet:

① → Any Dir. dist. can be represented as a normalized set of indep. Gamma Random Var.

$$Dir(\alpha_1, \ldots, \alpha_m) \equiv \left(\frac{Gam(\alpha_1)}{Gam(\alpha_1) + \cdots + Gam(\alpha_m)}, \ldots, \frac{Gam(\alpha_m)}{Gam(\alpha_1) + \cdots + Gam(\alpha_m)} \right)$$

② → Sum of Gamma RVs (with common scale parameters) is also a Gamma R.V.

$$Gam(\alpha_1 + \alpha_2) = Gam(\alpha_1) + Gam(\alpha_2)$$

$\cancel{Gam(\alpha_1 + \alpha_2 + \cdots + \alpha_m)}$

⇒ The aggregation of any subset of Dir. variables yield a Dir, with corresponding agg. of parameters.

$$q = (\underbrace{q_1, \ldots, q_i} + \underbrace{q_{i+1}, \ldots q_m}) \sim Dir(\underbrace{\alpha_1, \ldots \alpha_i} + \underbrace{\alpha_{i+1}, \ldots \alpha_m})$$

$$q = (q_1 + q_2 + \cdots + q_i, \; q_{i+1} + \cdots + q_m) \sim Dir(\alpha_1 + \cdots + \alpha_i, \; \alpha_{i+1} + \cdots + \alpha_m)$$

NB Methods

- Dirichlet process / Chinese Restaurant process
  Latent class models
  (often used in clustering)
- Beta process / Indian buffet process
  (latent feature models)
- Gaussian process
  (Regression)

---

* Gaussian process:

Defines a distribution $p(f)$ over functions f where f is a function mapping some input space $\mathcal{X}$ to R $\quad f: \mathcal{X} \to R$

($\mathcal{X}$ could be infinite dimentional quantity.
(e.g. $\mathcal{X} = R$)

Let $f = (f(x_1), f(x_2) --, f(x_n))$ n-dim. vector of function values evaluated at n points $x_i \in \mathcal{X} \Rightarrow f$ is a R.V.

then: $p(f)$ is a Gaussian process if for any finite subset $\{x_1, —, x_n\} \subset \mathcal{X}$ the marginal dist. over that finite subset $p(f)$ has a multivariate Gaussian dist.

GPs are parametrized by a mean function $\mu(x)$ and cov. function $C(x, x')$

$$p(f(x), f(x')) = N(\mu, \Sigma)$$

$$\mu = \begin{bmatrix} \mu(x) \\ \mu(x') \end{bmatrix}, \quad \Sigma = \begin{bmatrix} C(x,x) & C(x,x') \\ C(x',x) & C(x',x') \end{bmatrix}$$

$$C(x_i, x_j) = v_0 \exp\left\{-\left(\frac{|x_i - x_j|}{\lambda}\right)^{\alpha}\right\} + v_1 + v_2\,\delta_{ij}$$

with parameters $(v_0, v_1, v_2, \lambda, \alpha)$

Dirichlet dist.

is a distribution over K-dimensional prob. simplex.

Let $\underline{p}$ be a K-dim vector s.t. $\forall j$

$$p_j \geq 0 \ \ \& \ \ \sum_{j=1}^{K} p_j = 1$$

$$p(\underline{p}|\alpha) = \mathrm{Dir}(\alpha_1, \alpha_2, \ldots \alpha_K) \triangleq \frac{\Gamma(\sum_j \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{K} p_j^{\alpha_j - 1}$$

normalization const.

$$E(p_j) = \frac{\alpha_j}{\sum_k \alpha_k}$$

Dir is conjugate to multinomial dist.

Let $c|p \sim \text{multinomial}(\cdot|p)$

$p(c=j|p) = p_j$ Then posterior is Dirichlet

$$p(p|c=j, \alpha) = \frac{p(c=j|p)\,p(p|\alpha)}{p(c=j|\alpha)} = \mathrm{Dir}(\alpha')$$

$\alpha'_j = \alpha_j + 1 \quad \forall l \neq j$

$\alpha'_l = \alpha_l$

n integer

$\Gamma(n) = (n-1)!$

$p(f) \sim$ Gauss. process (GP)

paramtrized by a mean function $\mu(x)$

and cov. function $C(x, x')$

$$p(f(x), f(x')) = N(\mu, \Sigma)$$

where

$$\mu = \begin{bmatrix} \mu(x) \\ \mu(x') \end{bmatrix} ; \quad \Sigma = \begin{bmatrix} C(x,x) & C(x,x') \\ C(x',x) & C(x',x') \end{bmatrix}$$

for $p(f(x_1), \ldots f(x_n)) \Rightarrow \mu$ is $n \times 1$ vector & $\Sigma_{n \times n}$ matrix.

$$C(x_i, x_j) = V_0 \exp\left\{ -\left(\frac{|x_i - x_j|}{\lambda} \right)^{\alpha} \right\} + V_1 + V_2\, \delta_{ij}$$

with paramters $(V_0, V_1, V_2, \lambda, \alpha)$

Once the mean & cov. functions are defined then everything can be derived easily from MV Gaussian

How to use GP for nonlinear regression:

observing Data set $D = \{ (x_i, y_i)_{i=1}^{n} \} = (X, Y)$

<u>Model</u>

$$y_i = f(x_i) + \epsilon_i$$

$$f \sim GP(\cdot \mid \underline{0}, \underline{C})$$

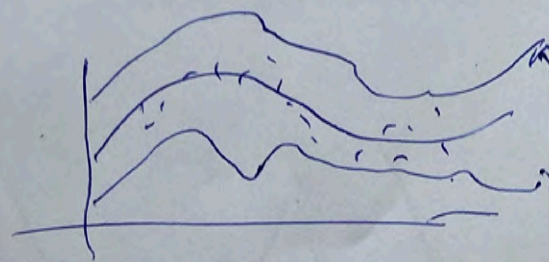
$$\epsilon_i \sim N(\cdot \mid 0, \delta^2)$$

prior on f is a GP, likelihood is Gaussian $\Big\} \Rightarrow$ posterior on f is also GP

$$p(y'|x';D) = \int p(y'|x',f,D)\,p(f|D)\,df$$

prediction

compute marginal likelihood to tune cov. function $\Rightarrow$ $p(Y|X) = \int p(Y|f,X)\,p(f)\,df$



prediction w 2 standard deviation

Consider Linear regression with input x_i and output Y_i ; $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Linear regression with K basis functions

$$Y_i = \sum_{k=1}^{K} \beta_k \phi_k(x_i) + \epsilon_i$$

Bayesian linear regression with basis function

$$\beta_k \sim N(\cdot|0, \lambda_k) \text{ (indep. of } \beta_l \text{ ; } \forall l \neq k) \text{ ; } \epsilon_i \sim N(\cdot|0, \delta^2)$$

Integrating coef. β_j we find:

$$E[y_i] = 0 \qquad \text{Cov.}(y_i, y_j) = C_{ij} \triangleq \sum_k \lambda_k \phi_k(x_i)\phi_k(x_j) + \delta_{ij}\,\delta^2$$

$\{y_i\}$ This is a Gaussian process with cov. function $C(x_i, x_j) = C_{ij}$

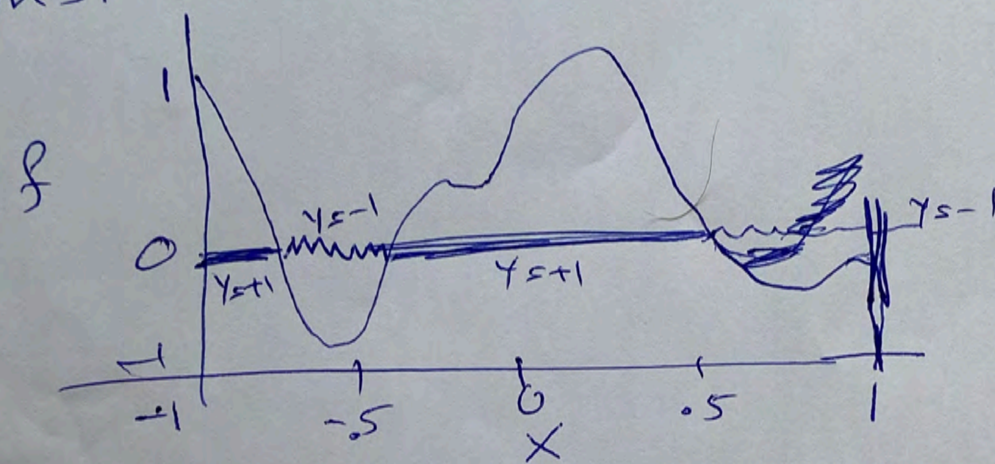
~~y_i Gaussian process with cov. function~~

this GP has finite # of basis functions (K)

Using GP for Classification
(2-class problem) — Binary

Given a data set $D=\{(x_i, y_i)\}_{i=1}^{n}$
with binary class labels $y_i \in \{-1,+1\}$

infer class label probabilities at new points.



Relate $f(x_i)$ to Class probabilities

$$p(y|f) = \begin{cases} \frac{1}{1+\exp(-yf)} & \rightarrow \text{Sigmoid (logistic)} \\ \Phi(yf) & \rightarrow \text{Cumulative normal (probit)} \\ H(yf) & \rightarrow \text{threshold} \\ \epsilon + (1-2\epsilon) H(yf) & \rightarrow \text{robust threshold} \end{cases}$$