Gaussian process classifiers :

Recall logistic regression :

$$p(y=+1 \mid x, w) = \sigma(x^T w)$$

$$p(y_i \mid x_i, w) = \sigma(y_i f_i)$$

often $\frac{1}{1+\exp(-w'x)}$

for LR, other sigmoids possib

Now write $p(w) = N(0, \Sigma_p)$

then $\log p(w \mid X, y) = -\frac{1}{2} w' \Sigma_p w + \sum_i \log$

$y_i f_i$

+ Const.

Our version so far is <u>maximum likelihood</u>

$\hat{=}$ choose $w = \text{argmax} \sum \log \sigma(y_i f_i)$

penalized LR is <u>maximum a posteriori</u>

$\hat{=}$ choose $w = \text{argmax} \left[\sum_i \log \sigma(y_i f_i) - \frac{1}{2} w' \Sigma_p w \right.$

Alternative:

$$p(y_* = 1 \mid x_*, D) = \int p(y_* = 1 \mid w, x_*)\, p(w \mid D]\, dw$$

i.e. average the prediction.

Gaussian process classification:

- replace w's with a GP
- notice we don't know $f(x)$ at the training points — nuisance parameters.
- classifying a point x_*

1): find f_*

$$p(f_* \mid X, y, x_*) = \int p(f_* \mid X, x_*, f) \cdot p(f \mid X, y)\, df$$

2): prediction

$$\pi_* = p(y_* = 1 \mid X, y, x_*) = \int p(y_* = 1 \mid X, y, x_*, f_*) \cdot p(f_* \mid X, y,$$

unknown

obs	f_i	obs
y_i	?	x_i
1	?	. . .
0	?	. . .
:		etc
.		

Notice $p(f^*/X, x_*, f)$ is straightforward from linear regression

$$N\Big(K(x_*, X)[K(X,X)]^{-1} f,$$

$$K(x_*, x_*) - K(x_*, x_*)[K(X,X)]^{-1} K(X, x_*) \Big)$$

but we need

$$p(f/X, y).$$

Laplace approx:

- approx $p(f/X,y)$ with a Gaussian
- mean is at argmax $p(f/X,y) = f_0$
- covar is $-H_f \log p(f/X,y)\Big|_{f=f_0}$

$$p(f/X,y) \propto p(f/X) \cdot p(y/f)$$

↑ GP prior ; ↑ from sigmoid

Now to find $f_0 = \text{argmax}\ p(f/X,y)$

$$\psi(f) \triangleq \log p(y/f) + \log p(f/X)$$

$$= \log p(y/f) - \frac{1}{2} f' K^{-1} f - \frac{1}{2} \log |K| - \frac{n}{2} \log 2\pi$$

$$\nabla_f \psi = \bullet \cdot \nabla_f \log p(y/f) - K^{-1} f$$

$$H_f \psi = H_f \log p(y/f) - K^{-1} = -W - K^{-1}$$

↑ this will be diagonal

Note $K = K[X,X]$

work with logistic

$$p(y_i | f_i) = -\log(1 + \exp(-y_i f_i))$$

$$\frac{\partial}{\partial f_i} \log p(y_i | f_i) = t_i - \pi_i$$

$$\left(\frac{1 + y_i}{2}\right) \qquad p(y_i = 1 | f_i) = -\log(1 + ex$$

$$\frac{\partial^2}{\partial f_i} \log p(y_i | f_i) = -\pi_i [1 - \pi_i]$$

Notice that at f_0,

$$\underline{f}_0 = K[\nabla \log p(y | \underline{f}_0)] \quad \text{(from gradien}$$

Use newton's method to find $\underline{f}_0$

then

$$p(f | X, y) \approx q(f | X, y) = N(f_0,$$

$$(K^{-1} + W)^{-1}$$

in what sense?

- integration

Prediction:

want posterior mean for f_*.

$$E[f_* | X, y, X_*] = \int \left[\int f_* \, p(f_* | X, y, X_*, f) \, df_* \right] \times p(f | X, y, X_*) \, df$$

$$= \int E[f_* | X, y, X_*, f] \cdot p(f | X, y, X_*) \, d$$

now approx p with q from above

$$E_q[f_* | X, y, X_*] = K[X_*, X] \left[K(X, X) \right]^{-1} f_0$$

Variance:

$$E\left[\left(f_* - E(f_*) \right)^2 \right]$$

$$= E\left[\left\{ \left(f_* - E(f_* | f) \right) + \left(E(f_* | f) - E(f_*) \right) \right\}^2 \right]$$

↑ not a function of f

so these two are indep

So:

$$\text{Var} = \underset{P(f_* \mid X, x_*, f)}{E}\left[\left(f_* - E(f_* \mid X, x_*, f)\right)^2\right]$$

$$+ \underset{q}{E}\left[\left(E(f_* \mid X, x_*, f) - E(f_* \mid X, y, x_*\right)\right.$$

- First term from GP regression

$$K(x_*, x_*) - K(x_*, X)\left[K(X, X)\right]^{-1} K(X, x_*)$$

- Second term from Gaussian approx to q

$$K(x_*, X)\, K(X, X)^{-1}\left(K^{-1} + W\right)^{-1} K[X, X]^{-1} K($$

- now our predictions are

$$\tilde{\pi}_* = E_q\left[\pi_* \mid X, y, x_*\right]$$

$$= \int \sigma(f_*^*)\, q(f_*)\, df_*$$

Assumed density filtering:

- method to deal with nasty poster

- $p(x/\theta) = (1-\omega) N(x; \theta, I) + \omega N($

signal ↑ ↑ clut

θ is Gaussian prior

$p(D, \theta)$ is quite nasty

multiple bumps

$p(D, \theta) = p(\theta) \prod_i p(x_i/\theta)$

- factor $p(D,\theta)$

- ie. eg. $t_0 = p(\theta)$, $t_i = p(x_i|$

- Now choose parametric

$$q(\theta) \sim N(m_\theta, v_\theta I)$$

now build

$q_{i+1}(\theta)$ by:

$$q_0(\theta) = t_0(\theta)$$

$$\hat{p}_{i+1}(\theta) = \frac{t_{i+1}(\theta)\, q_i(\theta)}{\int_\theta t_{i+1}(\theta)\, q_i(\theta)\, d\theta}$$

and $q_{i+1}(\theta)$ chosen to minimize

$$D\left(\hat{p}_{i+1}(\theta) \,\|\, q_{i+1}(\theta)\right)$$

- if the q_i are spherical gaussian

m_{θ_i}, $v_{i\theta}$

Alternative strategy: Expectation-propagation

must approx $p(f|X,y) = \frac{1}{Z} \cdot p(f|X) \cdot \prod_{i=1}^{n} p(y_i|f_i)$

posterior on function values

strategy: approx

$$p(y_i|f_i) \propto t_i(f_i|\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2)$$

$$\triangleq \tilde{Z}_i \, N(f_i|\tilde{\mu}_i, \tilde{\sigma}_i^2)$$

site params $\tilde{Z}_i, \tilde{\mu}_i, \tilde{\sigma}_i^2$

this gives

$$q(f|X,Y) \triangleq \frac{1}{Z_{EP}} p(f|X) \cdot \prod_i t_i$$

$$= N(\mu, \Sigma).$$

How to choose $\tilde{Z}_i, \mu_i, \sigma_i^2$?

- Start from some approximation
- leave out t_i term
  → cavity dist q_{-i}
- now choose site pars to best approx $q_{-i}\, p(y_i | f_i)$

Iterate:

cavity dist

$$q_{-i}(f_i) = \int \left[p(f|X) \prod_{j \neq i} t_j(f_j) \right] df_{-i}$$

this is gaussian, so marginal is easy.

$$q_{-i}(f_i) = N(\mu_{-i}, \sigma_{-i}^2)$$