Gaussian process classification model

- 'training' data $D = \{(\overset{\in \mathbb{R}}{x_1}, t_1), (x_2, t_2), \ldots, (x_N, t_N)\} \in [\mathbb{R} \times \{0, 1\}]^N$
  class label $\in \{0,1\}$
- 'test' data: $\tilde{x}$
- model: $y(x) \sim \text{GP}$ with $m(x) = 0$, kernel $k(x, x')$ (latent process)
  observation: $t_i = \text{Bernoulli}(\underbrace{\sigma(\overbrace{y(x_i)}^{y_i})}))$ link fn (sigmoid) : $\mathbb{R} \to [0,1]$
  (i.e., $p(t|y_i) = \sigma(y_i)^t(1 - \sigma(y_i))^{1-t}$)

$$t(x) = \begin{cases} 0 & \text{wp } \frac{1}{1+e^{y(x)}} = \sigma(-y(x)) \\ 1 & \text{wp } \frac{e^{y(x)}}{1+e^{y(x)}} = \sigma(y(x)) \end{cases}$$

- prior: with K_D, k, c as in GP regression
  on $y(x)$

$$(y_1, y_2, \ldots, y_N, \tilde{y}) \sim \mathcal{N}\left(0, \begin{bmatrix} K_D & k \\ k^\mathsf{T} & c \end{bmatrix}\right)$$

$K_D = \{k(x_i, x_j)\}$, $k = \{k(x_i, \tilde{x})\}$, $c = k(\tilde{x}, \tilde{x})$

- posterior: how do we compute $p(\tilde{y}|D)$?

posterior

- 'training' data $D = \{(x_1, t_1), (x_2, t_2), \ldots, (x_N, t_N)\} \in [\mathbb{R} \times \{0, 1\}]^N$
- model: $y(x) \sim$ GP with $m(x) = 0, k(x, x'),\ t_i = \text{Bernoulli}(\sigma(y(x_i)))$
- likelihood given $y_i = y(x_i)$

$$\log p(t|y(x)) = t^T y + \sum_{i=1}^{N} \log(1 + e^{y_i})$$

- negative log of posterior $-\log p(y|t)$ ← approx this as a quadratic

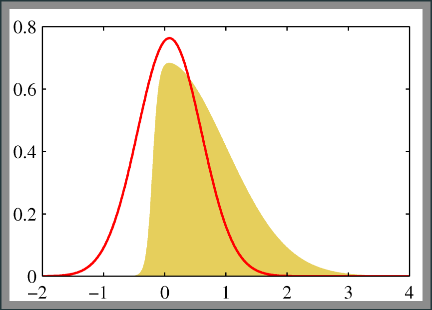
$$\psi(y) = \underbrace{\frac{1}{2} y^T K^{-1} y + \frac{1}{2} \log|K|}_{-\text{log of prior}} - \underbrace{\left(t^T y + \sum_{i=1}^{N} \log(1 + e^{y_i})\right)}_{\text{-ive log of likelihood (prev page)}} + \underbrace{\text{const}}_{\text{normalization}}$$

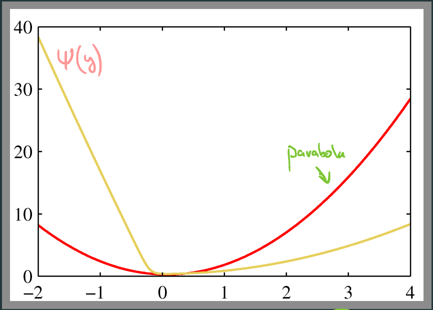
prior $p(y) = (2\pi)^{-N/2} (|K|)^{-1/2} \exp\left(-\frac{(y-m)^T K^{-1} (y-m)}{2}\right)$

the Laplace approximation

approximate posterior as a multivariate Gaussian

i.e. 'match the mode and the Hessian'





Want $p(y|D) \sim N(\tilde{\mu}, \tilde{\Sigma}) \Rightarrow -\log p(y|D) \approx \frac{1}{2}(y-\tilde{\mu})^T \tilde{\Sigma}^{-1}(y-\tilde{\mu})$
$+ \frac{1}{2}\log|\tilde{\Sigma}|$

Laplace approx - set $\tilde{\mu} = \nabla \psi(y)\big|_{\tilde{\mu}} = 0$, $\tilde{\Sigma} = \nabla\nabla\psi(y)$

Laplace approximation for GP classification

Final output

$$p(y|D) \sim \mathcal{N}(y^*, H^{-1})$$

where $K^{-1} y^* = t - \sigma_N(y^*)$

$$H^{-1} = K^{-1} + W^*$$

$\tilde{y}(\tilde{x}) \sim \mathcal{N}(\cdot, \cdot)$

$$\mathbb{E}[\tilde{y}(\tilde{x})|D] = k^T K^{-1} y^* = k^T (t - \sigma_N(y^*))$$

$$\mathrm{Var}(\tilde{y}(\tilde{x})|D) = c - k^T H k = c - k^T (K^{-1} + W^*)^{-1} k$$

and $t(\tilde{x}) \sim \mathrm{Bernoulli}(\sigma(\tilde{y}))$

Laplace approximation for GP classification

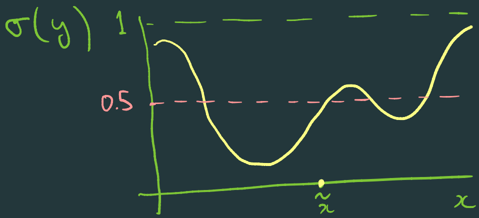
Q: Given $\tilde{x}$, what do we want to predict?

A: Depends on loss fn ...

'true class label' for $\tilde{x}$

Typical setting (L_0 loss): $L(t(\tilde{x})) = \mathbb{E}\left[\mathbb{1}\{t(\tilde{x}) \neq \tilde{t}\}\right]$

output of classifier



- 'Bayes classifier'
- $L(t(\tilde{x})) = p(\tilde{t}=1)\cdot \overbrace{\mathbb{1}\{t(\tilde{x})=0\}}^{1-\mathbb{1}\{t(\tilde{x})=1\}}$

$\qquad + \; p(\tilde{t}=0)\cdot \mathbb{1}\{t(\tilde{x})=1\}$

$= \underbrace{p(\hat{t}=1)}_{\text{can't control}} + \mathbb{1}\{t(\tilde{x})=0\}\underbrace{\left(p(\hat{t}=1) - p(\hat{t}=0)\right)}_{\substack{\text{If +ive, set } t(\tilde{x})=1 \\ \text{else set } t(\tilde{x})=0}}$

i.e. - Bayes classifier is the MAP estimator for $t(\tilde{x})$

Laplace approximation for GP classification

For Bayes classifier, need to compute

$$\mathbb{P}[t(\tilde{x})=1 \mid D] = p(\tilde{t} \mid D) = \int \sigma(\tilde{y})\, p(\tilde{y} \mid D)\, d\tilde{y}$$

$\frac{e^{\tilde{y}}}{1+e^{\tilde{y}}}$ complicated fn :(

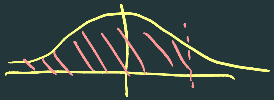
$\mathcal{N}(\mu_D, \Sigma_D)$

known in closed form! (Sec 4.5 of Bishop)

How can we evaluate this?

1) Variational Approx – replace $\sigma(\tilde{y})$ by 'probit link fn' $\Phi(\tilde{y})$ = 'tail prob of the Gaussian'

2) Monte Carlo simulation



Laplace approximation: model selection

Idea - Maximize marginal likelihood $p(t|\theta)$ (ie, min $-\log p(t|\theta)$)

(Bayesian Occam's Razor)

θ = hyper params

$$p(t|\theta, x) = \frac{\overbrace{p(y|\theta)}^{\text{Prior}}\ \overbrace{p(t|y)}^{\text{likelihood}}}{\underbrace{p(y|\theta, D)}_{\text{posterior}}}$$

$\tilde{q}(t|\theta, x)$

$\tilde{L}(y) \equiv N(y^*, H^{-1})$

Laplace approx

$N(\tilde{\mu}_D, \tilde{\Sigma}_D)$

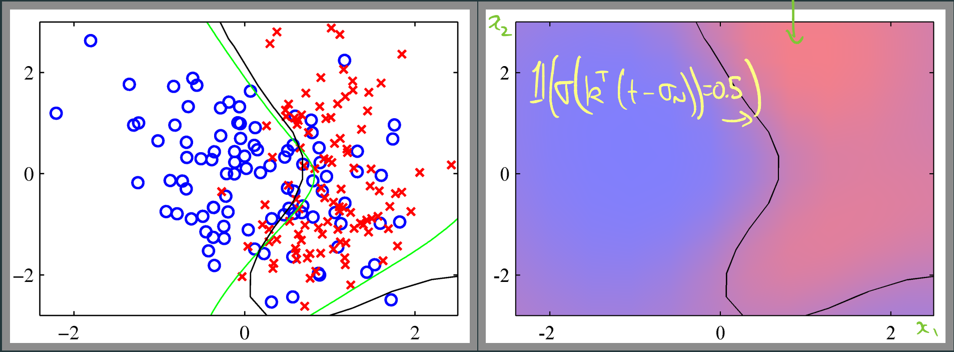
- posterior, likelihood not available in closed form — Use Laplace approx instead

$$-\log\left(\tilde{q}(t|\theta, x)\right) = \frac{1}{2} y^{*T} K^{-1} y^* + \frac{1}{2}\log|K| + \frac{1}{2}\log|K^{-1} + W| - \underbrace{\log\left[p(t|y^*)\right]}_{\sigma(y^*)^t (1-\sigma(y^*))^{1-t}}$$

Laplace approx for latent vars

classification using GPs: decision boundaries

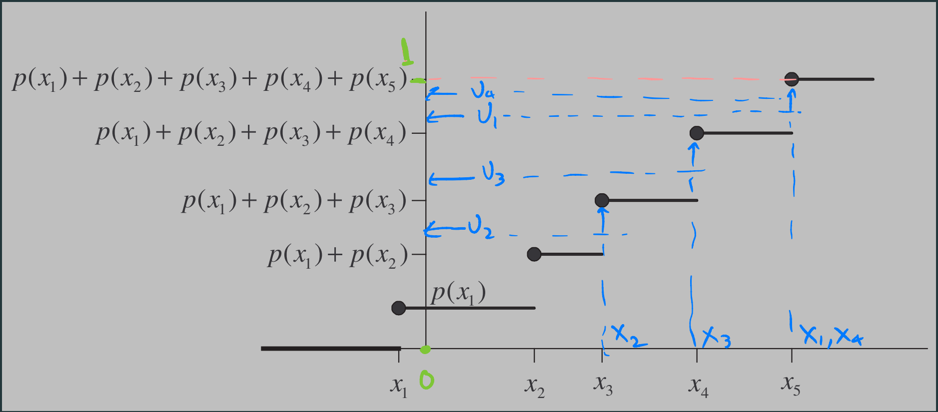
$\sigma(y(x))$ $x \in \mathbb{R}^2$



basic monte carlo techniques

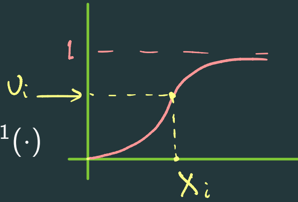
warmup: simulating discrete rv

X takes values $x_1 \leq x_2 \leq \ldots \leq x_5$, $\mathbb{P}[X = x_i] = p(x_i)$



the inversion method

X continuous r.v. with pdf f and c.d.f. $F(\cdot)$



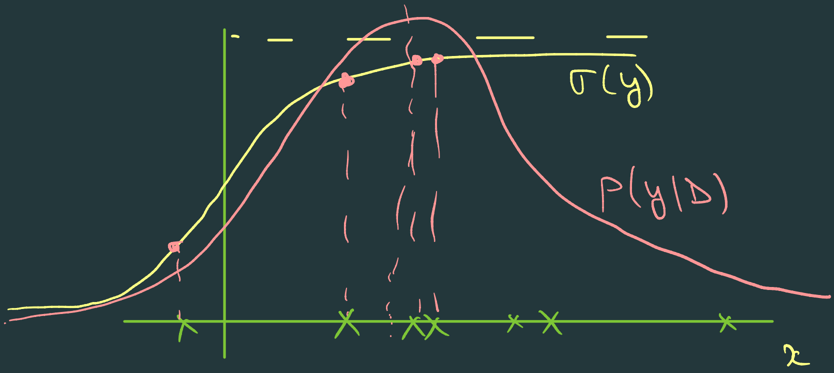
- want to generate samples of X.
- $F(\cdot)$ non-decreasing $\implies$ can define inverse $F^{-1}(\cdot)$
- $F(x) = u \iff F^{-1}(u) = x$

inversion method

given desired cdf F (continuous, increasing), generate sample $X_0 \sim F$ as:

1. generate $U \sim U[0, 1]$.
2. return $X_o = F^{-1}(U)$.

Application - Compute integrals



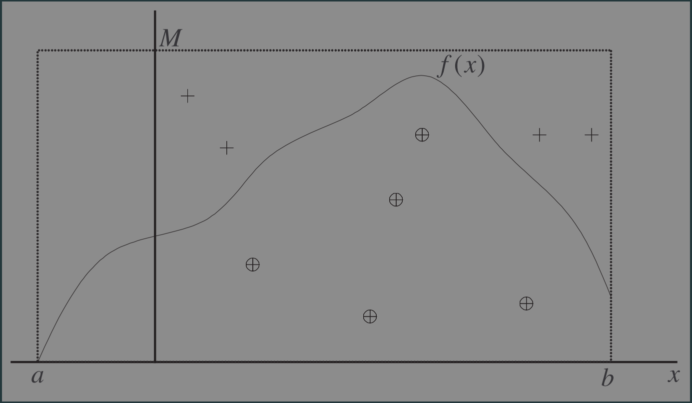
$$I = \mathbb{E}_{x \sim F}\left[\sigma(x)\right] \approx \frac{1}{N} \sum_{i=1}^{N} \sigma(x_i)$$

rejection sampling

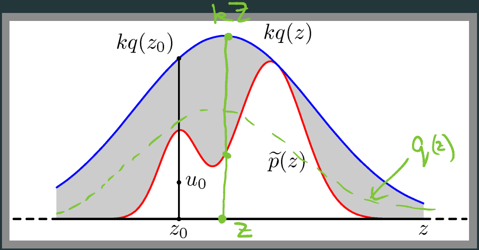
want samples of a rv $X \in [a, b]$, with pdf $f(x) \leq M$

rejection sampling

1. Generate $U_1, U_2 \sim U[0, 1]$, and set $Z_1 = a + (b - a)U_1$, $Z_2 = MU_2$
2. if $Z_2 \leq f(Z_1)$, return $X_o = Z_1$; else, reject and repeat



generalized rejection sampling



- Given a 'sampler' $Z \sim Q$
- Want samples $X \sim P$

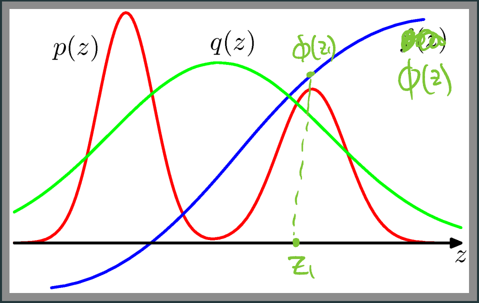
- find k s.t. $kq(x) \geq p(x) \; \forall x$ (ie, $k \geq \max \frac{p(x)}{q(x)}$)
- Generate $Z \sim Q$
- Accept (ie set $X = Z$) w.p $\frac{p(Z)}{kq(Z)}$, else repeat

importance sampling (for estimating integrals)

- given function $\phi(\cdot)$, want $\mathbb{E}[\phi(X)]$ where $X \sim P$
- can generate samples $Z \sim Q$

importance sampling

1. generate $Z_1, Z_2, \dots, Z_L \sim Q$
2. compute $\mathbb{E}[\phi(X)] = \frac{1}{L}\sum_{i=1}^{L} w_i \phi(Z_i)$, where $w_i = p(Z_i)/q(Z_i)$



$$\mathbb{E}_{x\sim p}[\phi(x)] = \int \phi(x)\, p(x)\, dx$$

$$= \int \phi(x) \left(\frac{p(x)}{q(x)}\right) q(x)\, dx$$

$$= \mathbb{E}_{z\sim Q}[\phi(z)\, w(z)]$$

$p(z)/q(z)$

$$= \frac{1}{L}\sum_{i=1}^{L} w_i \phi(z_i)$$