Announcements

Hi

Matlab Grader homework, emailed Thursday,
1 (of 9) homeworks Due 21 April, Binary graded.
2 this week

Jupyter homework?: translate matlab to Jupiter, TA Harshul h6gupta@eng.ucsd.edu or me
I would like this to happen.

"GPU" homework. NOAA climate data in Jupyter on the datahub.ucsd.edu, 15 April.

Projects: Any computer language

Podcast might work eventually.

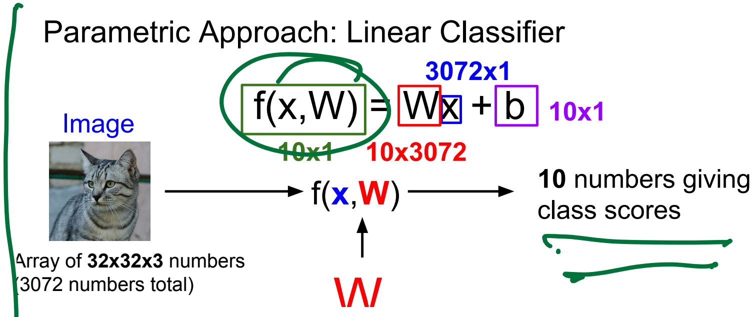
Today:
- Stanford CNN
- Gaussian, Bishop 2.3
- Gaussian Process 6.4
- Linear regression 3.0-3.2

Wednesday 10 April
Stanford CNN, Linear models for regression 3, Applications of Gaussian processes.

Bayes and Softmax (Bishop p. 198)

- Bayes:

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



- Classification of N classes:

$$p(\mathcal{C}_n|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_n)p(\mathcal{C}_n)}{\sum_{k=1}^{N} p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}$$

$$= \frac{\exp(a_n)}{\sum_{k=1}^{N} \exp(a_k)}$$

≤ 1

≥ 0

$\sum 1$

with

$$a_n = \ln\left(p(\mathbf{x}|\mathcal{C}_n)p(\mathcal{C}_n)\right)$$

Softmax to Logistic Regression (Bishop p. 198)

$$
\begin{aligned}
p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{\sum_{k=1}^{2} p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)} \\
&= \frac{\exp(a_1)}{\sum_{k=1}^{2}\exp(a_k)} = \frac{1}{1+\exp(-a)}
\end{aligned}
$$

with

$$
a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}
$$

Softmax with Gaussian(Bishop p. 198)

$$p(\mathcal{C}_n|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_n)p(\mathcal{C}_n)}{\sum_{k=1}^{N} p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}$$

$$= \frac{\exp(a_n)}{\sum_{k=1}^{N} \exp(a_k)}$$

with

$$a_n = \ln\left(p(\mathbf{x}|\mathcal{C}_n)p(\mathcal{C}_n)\right)$$

Assuming $\mathbf{x}$ is Gaussian $\mathcal{N}(\mu_n, \mathbf{\Sigma})$

$$a_n = \mathbf{w}_n^T\mathbf{x} + w_0$$

$$\mathbf{w}_n = \mathbf{\Sigma}^{-1}\mu_n$$

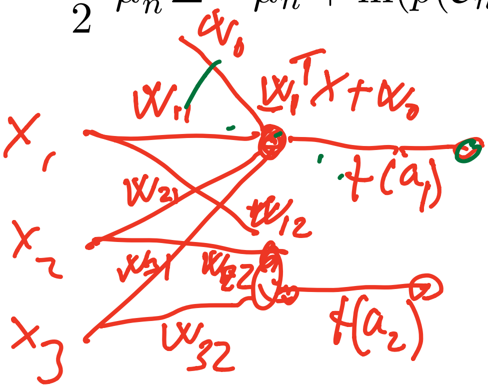
$$w_0 = \frac{-1}{2}\mu_n^T\mathbf{\Sigma}^{-1}\mu_n + \ln(p(\mathcal{C}_n))$$

$$x \in N(\mu_n, \Sigma)$$

$$\propto e^{-\frac{1}{2}(x-\mu_n)^T\Sigma^{-1}(x-\mu_n)}$$

$$-\ln(p(x|C_n)) \sim \cancel{\frac{1}{2}x^T\Sigma^{-1}x} + \frac{1}{2}\mu^T\Sigma^{-1}\mu \underbrace{- \mu_n^T\Sigma^{-1}}_{w_n}x$$

$$w_0 = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \ln C_n$$

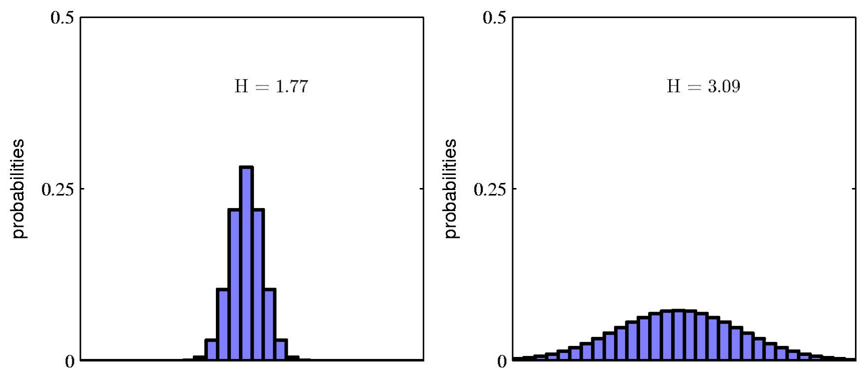


Entropy 1.6

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning



The Kullback-Leibler Divergence

P true distribution, q is approximating distribution

$$\begin{aligned}
\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\, \mathrm{d}\mathbf{x}\right) \\
&= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x}
\end{aligned}$$

not a distance measure

$$\mathrm{KL}(p\|q) \geqslant 0 \qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

KL homework

- Support of P and Q = > “only >0” don’t use isnan isinf

- After you pass. Take your time to clean up. Get close to 50

Lecture 3

- Homework

- Pod-cast lecture on-line

- Next lectures:
  - I posted a rough plan.
  - It is flexible though so please come with suggestions

Bayes for linear model

$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$ $\quad$ $\boldsymbol{n} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{C}_n)$ $\quad$ $\mathbf{y} \sim \mathrm{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{C}_n)$ $\quad$ prior: $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{C}_x)$

$p(\boldsymbol{x}|\boldsymbol{y}) \sim p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) \sim N(\boldsymbol{x}_p, \boldsymbol{C}_p)$

mean $\quad \boldsymbol{x}_p = \boldsymbol{C}_p \boldsymbol{A}^T \boldsymbol{C}_n^{-1} \boldsymbol{y}$

Covariance $\quad \boldsymbol{C}_p^{-1} = \boldsymbol{A}^T \boldsymbol{C}_n^{-1} \boldsymbol{A} + \boldsymbol{C}_x^{-1}$

$$\propto e^{-\frac{1}{2}(x - x_p)^T C_p^{-1}(x - x_p)}$$

$$\propto e^{-\frac{1}{2}(Ax - y)^T C_n^{-1}(Ax - \mu)} e^{-\frac{1}{2} x C_x^{-1} x}$$

$$\propto e^{-\frac{1}{2}[\underbrace{x^T A^T C_n^{-1} A x + x^T C_x x}_{x^T C_p^{-1} x} - \underbrace{x^T A^T C_n^{-1} \mu}_{x^T C_p x_p}}$$

Bayes' Theorem for Gaussian Variables

- Given

$$
\begin{aligned}
p(\mathbf{x}) &= \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}\right) \\
p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\mathbf{y}|\mathbf{A}\mathbf{x}+\mathbf{b}, \mathbf{L}^{-1}\right)
\end{aligned}
$$

- we have

$$
\begin{aligned}
p(\mathbf{y}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}+\mathbf{b}, \mathbf{L}^{-1}+\mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathrm{T}}) \\
p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^{\mathrm{T}}\mathbf{L}(\mathbf{y}-\mathbf{b})+\boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})
\end{aligned}
$$

- where

$$
\boldsymbol{\Sigma} = (\boldsymbol{\Lambda}+\mathbf{A}^{\mathrm{T}}\mathbf{L}\mathbf{A})^{-1}
$$

Sequential Estimation of mean (Bishop 2.3.5)

Contribution of the N^{th} data point, x_N

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathbf{x}_n \\
&= \frac{1}{N}\mathbf{x}_N + \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} \\
&= \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}\left(\mathbf{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}\right)
\end{aligned}
$$

correction given x_N

correction weight

old estimate

Bayesian Inference for the Gaussian (Bishop2.3.6)

Assume σ^2 is known. Given i.i.d. data $\mathbf{x} = \{x_1, \ldots, x_N\}$ the likelihood function for μ is given by

$$L(x) \qquad p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

- This has a Gaussian shape as a function of μ (but it is *not* a distribution over μ).

Bayesian Inference for the Gaussian (Bishop2.3.6)

- Combined with a **Gaussian prior over μ,** $$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$
- this gives the posterior

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

$$\frac{(\mu-\mu_N)^2}{2\sigma_N^2} = \frac{1}{2\sigma^2}\sum(x-\mu)^2 + \frac{(\mu-\mu_0)^2}{2\sigma_0^2}$$

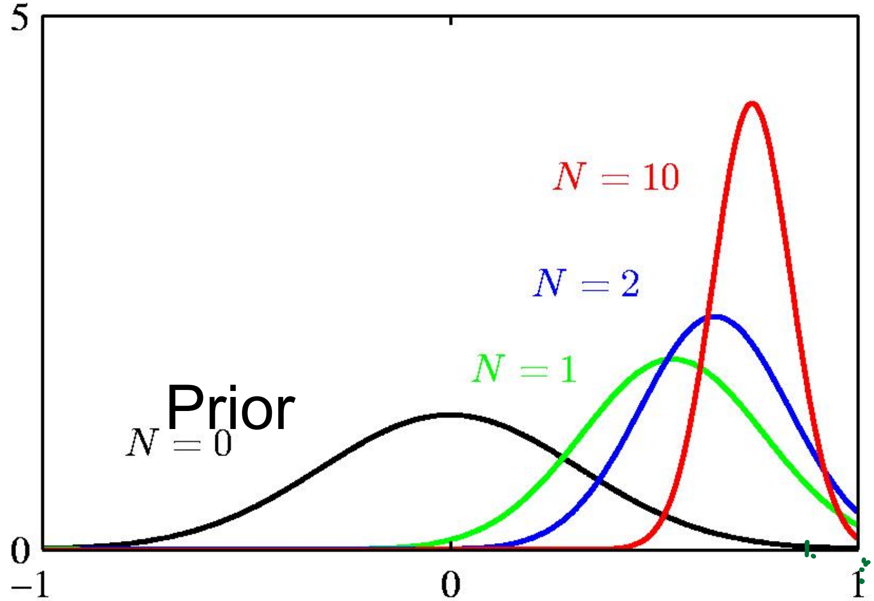
	$N = 0$	$N \to \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

σ^2

Bayesian Inference for the Gaussian (3)

- Example: for N = 0, 1, 2 and 10.

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$



Bayesian Inference for the Gaussian (4)

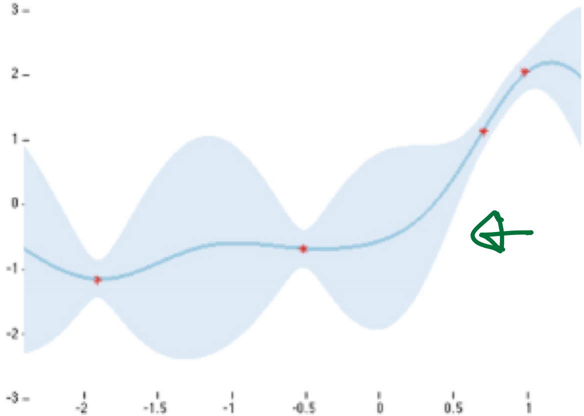
Sequential Estimation

$$
\begin{aligned}
p(\mu|\mathbf{x}) &\propto p(\mu)p(\mathbf{x}|\mu) \\
&= \underbrace{\left[p(\mu) \prod_{n=1}^{N-1} p(x_n|\mu) \right]} p(x_N|\mu) \\
&\propto \mathcal{N}\left(\mu|\mu_{N-1}, \sigma^2_{N-1}\right) p(x_N|\mu)
\end{aligned}
$$

The posterior obtained after observing N-1 data points becomes the prior when we observe the N th data point.

Conjugate prior: posterior and prior are in the same family. The **prior** is called a **conjugate prior** for the likelihood function.

Gaussian Process (Bishop 6.4, Murphy15)



$$t_n = y_n + \epsilon_n$$

$$\underline{f(\mathbf{x})} \sim GP(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$$

This is what a Gaussian process *posterior* looks like with 4 data points and a squared exponential covariance function. The bold blue line is the predictive mean, while the light blue shade is the predictive uncertainty (2 standard deviations). The model uncertainty is small near the data, and increases as we move away from the data points.

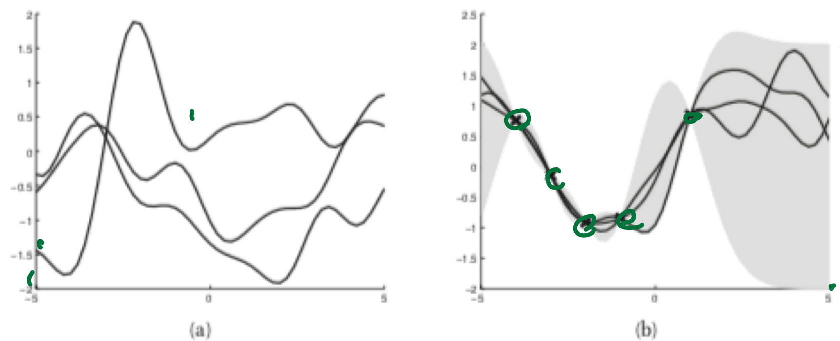


Figure 15.2 Left: some functions sampled from a GP prior with SE kernel. Right: some samples from a GP posterior, after conditioning on 5 noise-free observations. The shaded area represents $\mathbb{E}[f(\mathbf{x})] \pm 2\text{std}(f(\mathbf{x}))$. Based on Figure 2.2 of (Rasmussen and Williams 2006). Figure generated by `gprDemoNoiseFree`.

Gaussian Process (Murphy ch15)

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$$

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ \kappa(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))^T\right] \end{aligned}$$

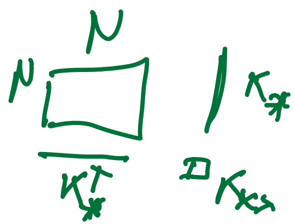
$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K})$$

$$K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \text{ and } \boldsymbol{\mu} = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_N))^T.$$

Training

$\mathcal{D} = \{(\mathbf{x}_i, f_i), i = 1 : N\}$, where $f_i = f(\mathbf{x}_i)$ is the noise-free

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$



$\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$ is $N \times N$, $\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*)$ is $N \times N_*$, and $\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*)$ is $N_* \times N_*$.

Gaussian Process (Murphy ch15)

$$\begin{pmatrix} \mathbf{f} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{pmatrix}, \begin{pmatrix} \mathbf{K} & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$

The conditional is Gaussian:

$$\begin{aligned} p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{f}) &= \mathcal{N}(\mathbf{f}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*) \\ \boldsymbol{\mu}_* &= \boldsymbol{\mu}(\mathbf{X}_*) + \mathbf{K}_*^T \mathbf{K}^{-1}(\mathbf{f} - \boldsymbol{\mu}(\mathbf{X})) \\ \boldsymbol{\Sigma}_* &= \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}^{-1} \mathbf{K}_* \end{aligned}$$

Common kernel is the squared exponential, RBF, Gaussian kernel

$$\kappa(x, x') = \sigma_f^2 \exp(-\frac{1}{2\ell^2}(x - x')^2)$$

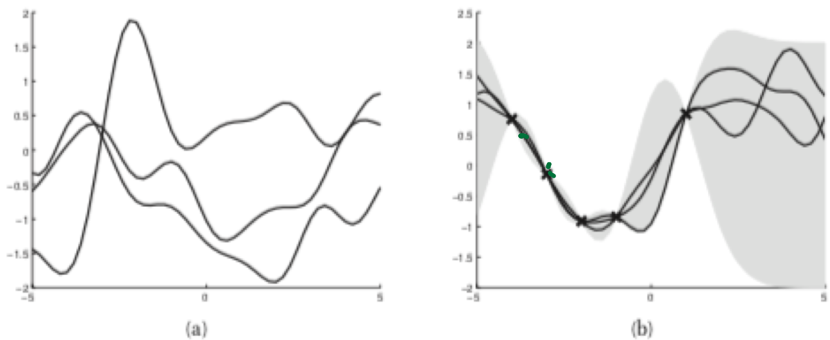


Figure 15.2 Left: some functions sampled from a GP prior with SE kernel. Right: some samples from a GP posterior, after conditioning on 5 noise-free observations. The shaded area represents $\mathbb{E}[f(\mathbf{x})] \pm 2\text{std}(f(\mathbf{x}))$. Based on Figure 2.2 of (Rasmussen and Williams 2006). Figure generated by `gprDemoNoiseFree`.