Piazza started

**Announcements**

**Matlab Grader homework**, email Friday,
2 (of 9) homeworks Due 21 April, Binary graded.

**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 language

**Podcast** might work eventually.

**Today:**
- Stanford CNN
- Bernoulli
- Gaussian 1.2
- Gaussian 2.3
- Decision theory 1.5
- Information theory 1.6

Monday
Stanford CNN, Linear models for regression 3
Non-parametric method

K-Nearest Neighbors

Instead of copying label from nearest neighbor, take *majority vote* from K closest points.
Interpreting a Linear Classifier

\[ f(x, W) = Wx + b \]

Array of 32x32x3 numbers (3072 numbers total)
Hard cases for a linear classifier

Class 1:
number of pixels > 0 odd

Class 2:
number of pixels > 0 even

Class 1:
1 <= L2 norm <= 2

Class 2:
Everything else

Class 1:
Three modes

Class 2:
Everything else
Coin estimate (Bishop 2.1)

- Binary variables $x=\{0,1\}$
  
  \[ p(x = 1|\mu) = \mu \]

- Bernoulli distributed
  
  \[ \text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x} \]
  
  \[ \mathbb{E}[x] = \mu \]
  
  \[ \text{var}[x] = \mu(1 - \mu). \]

- $N$ observations, Likelihood:
  
  \[ p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1 - \mu)^{1-x_n}. \]  

\[ \ell = \ln p(D|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln (1 - \mu)\}. \]  

- Max likelihood
  
  \[ \frac{\partial \ell}{\partial \mu} = \sum \frac{x_n}{\mu} + \frac{1-x_n}{1-\mu} \]

\[ \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \]
Coin estimate (Bishop 2.1)

- Bayes: $p(x|y) = p(y|x)p(x)$

- Conjugate prior

Bayes:

\[ a = 2 \]

\[ b = 2 \]

\[ \text{Beta}(\mu|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \]
ML MAP BAYES

- ML point estimate

- MAP point estimate (often in literature ML=MAP)

- Bayes => probability => From which all information can be obtained
  - MAP, median, error estimates
  - Further analysis as sequential
  - Disadvantage... not a point estimate.

Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a=2$, $b=2$, and the likelihood function, given by (2.9) with $N=m=1$, so that the posterior is given by a beta distribution with parameters $a=3$, $b=2$.

We see that this sequential approach to learning arises naturally when we adopt a Bayesian viewpoint. It is independent of the choice of prior and of the likelihood function and depends only on the assumption of i.i.d. data. Sequential methods make use of observations one at a time, or in small batches, and then discard them before the next observations are used. They can be used, for example, in real-time learning scenarios where a steady stream of data is arriving, and predictions must be made before all of the data is seen. Because they do not require the whole data set to be stored or loaded into memory, sequential methods are also useful for large data sets.

Maximum likelihood methods can also be cast into a sequential framework. Section 2.3.5 If our goal is to predict, as best we can, the outcome of the next trial, then we must evaluate the predictive distribution of $x$, given the observed data set $D$. From the sum and product rules of probability, this takes the form

$$p(x=1|D) = \int_0^1 p(x=1|\mu)p(\mu|D)d\mu = E[\mu|D].$$

(2.19)

Using the result (2.18) for the posterior distribution $p(\mu|D)$, together with the result (2.15) for the mean of the beta distribution, we obtain

$$p(x=1|D) = \frac{m+a}{m+l+b}$$

(2.20)

which has a simple interpretation as the total fraction of observations (both real observations and fictitious prior observations) that correspond to $x=1$. Note that in the limit of an infinitely large data set $m, l \to \infty$ the result (2.20) reduces to the maximum likelihood result (2.8). As we shall see, it is a very general property that the Bayesian and maximum likelihood results will agree in the limit of an infinitely large data set.
Bayes Rule

\[
P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}
\]

- Bayes rule tells us how to do inference about hypotheses from data.
- Learning and prediction can be seen as forms of inference.
The Gaussian Distribution

\[ \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \]

\[ \mathcal{N}(x|\mu, \sigma^2) > 0 \]

\[ \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, dx = 1 \]

Gaussian Mean and Variance

\[ \mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, x \, dx = \mu \]

\[ \mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) \, x^2 \, dx = \mu^2 + \sigma^2 \]

\[ \text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \]
Gaussian Parameter Estimation

Likelihood function

\[ p(x|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) \]

\[ \frac{\partial L}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu) = 0 \]

Maximal (Log) Likelihood

\[ \ln p(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \]

\[ \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

\[ \sigma^2_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})^2 \]
Curve Fitting Re-visited, Bishop 1.2.5

$$y(x, w)$$

$$y(x_0, w)$$

$$p(t|x_0, w, \beta) = \mathcal{N}(t|y(x_0, w), \beta^{-1})$$

$$2\sigma$$
Maximum Likelihood

\[ p(t|x, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, w), \beta^{-1}) . \]  

(1.61)

As we did in the case of the simple Gaussian distribution earlier, it is convenient to maximize the logarithm of the likelihood function. Substituting for the form of the Gaussian distribution, given by (1.46), we obtain the log likelihood function in the form

\[ \ln p(t|x, w, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi). \]  

(1.62)

\[ \frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, w_{ML}) - t_n\}^2 . \]  

(1.63)

Giving estimates of \( W \) and beta, we can predict

\[ p(t|x, w_{ML}, \beta_{ML}) = \mathcal{N}(t | y(x, w_{ML}), \beta_{ML}^{-1}) . \]  

(1.64)
MAP: A Step towards Bayes 1.2.5

prior
\[ p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}w^Tw\right\} \]

\[ p(w|x, t, \alpha, \beta) \propto p(t|x, w, \beta)p(w|\alpha) \]
\[ \mathcal{N}(w^Tx, \beta^{-1}) \mathcal{N}(0, \alpha) \]
\[ -\ln(p(w|t)) = \frac{\beta}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)^2 + \frac{\alpha}{2} w^Tw + \text{const} \]

\[ \beta\tilde{E}(w) = \frac{\beta}{2} \sum_{n=1}^{N} \left(y(x_n, w) - t_n\right)^2 + \frac{\alpha}{2} w^Tw \]

Determine \( w_{\text{MAP}} \) by minimizing regularized sum-of-squares error, \( \tilde{E}(w) \).

Regularized sum of squares
\[ p(t|x, w_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, w_{ML}), \beta^{-1}_{ML}) \]
Parametric Distributions

Basic building blocks:
Need to determine $\theta$ given $\{x_1, \ldots, x_N\}$

Representation: $\theta^*$ or $p(\theta)$

Recall Curve Fitting

$$p(t|x, x, t) = \int p(t|x, w)p(w|x, t) \, dw$$

We focus on Gaussians!
The Gaussian Distribution

\[ \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \]

\[ \mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\} \]
Central Limit Theorem

- The distribution of the sum of $N$ i.i.d. random variables becomes increasingly Gaussian as $N$ grows.
- Example: $N$ uniform [0,1] random variables.
Geometry of the Multivariate Gaussian

\[ e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \]

\[ \Delta^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \]

\[ \Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} u_i u_i^T \]

\[ \Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} = y^T \Sigma_1 y \]

\[ y_i = u_i^T (x - \mu) \]
Moments of the Multivariate Gaussian (2)

\[ \mathbb{E}(\mathbf{x}) = \mathbf{\mu} \]

\[ \mathbb{E}[\mathbf{xx}^T] = \mathbf{\mu}\mathbf{\mu}^T + \mathbf{\Sigma} \]

\[ \text{cov}[\mathbf{x}] = \mathbb{E} \left[ (\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^T \right] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \gamma_1 \gamma_2 & \hdots \gamma_1 \gamma_D \\ \gamma_1 \gamma_2 & \sigma_2^2 & \hdots \gamma_2 \gamma_D \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_1 \gamma_D & \gamma_2 \gamma_D & \hdots & \sigma_D^2 \end{bmatrix} \]

A Gaussian requires \( D^*(D-1)/2 + D \) parameters. Often we use \( D + D \) or just \( D + 1 \) parameters.
Partitioned Conditionals and Marginals

\[ p(x_a | x_b) = \mathcal{N}(x_a | \mu_{a|b}, \Sigma_{a|b}) \]

\[ \Sigma_{a|b} = \Lambda_{aa}^{-1} = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba} \]

\[ \mu_{a|b} = \Sigma_{a|b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \} \]

\[ = \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (x_b - \mu_b) \]

\[ = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b) \]

\[ p(x_a) = \int p(x_a, x_b) \, dx_b \]

\[ = \mathcal{N}(x_a | \mu_a, \Sigma_{aa}) \]

\[ p(x_a | x_b = 0.7) \]

\[ p(x_a) \]
ML for the Gaussian (1) Bisphop 2.3.4

Given i.i.d. data $X = (x_1, \ldots, x_N)^T$, the log likelihood function is given by

$$\ln p(X|\mu, \Sigma) = -\frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T \Sigma^{-1} (x_n - \mu)$$

$$\mathcal{L} = -\ln p$$

$$= \frac{N}{2} \ln |\Sigma| + \frac{1}{2} \text{tr} \left( \sum \frac{1}{n} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

$$= \frac{N}{2} \ln |\Sigma| + \text{tr} (\sum \frac{1}{n} (x - \mu) (x - \mu)^T \Sigma^{-1})$$

$$= \frac{N}{2} \left[ \ln |\Sigma| + \text{tr} (\sum \frac{1}{n} \Sigma^{-1}) \right]$$

$$\Sigma = \sum x - \mu (x - \mu)^T$$

$$\frac{\partial \Sigma}{\partial \mu} = \Sigma^{-1} + \sum \Sigma^{-1} = 0$$

$$\Sigma = \sum x - \mu (x - \mu)^T$$

$$\text{tr}(AB\Sigma) = \text{tr}(CAB)$$

$$\frac{\partial}{\partial A} \ln |A| = (A^{-1})^T$$

$$\frac{\partial}{\partial A} \text{tr}(AB) = B^T$$

$$\frac{\partial}{\partial x} (A^{-1}) = -A^{-1} \frac{\partial A}{\partial x} A^{-1}$$
Maximum Likelihood for the Gaussian

- Set the derivative of the log likelihood function to zero,

\[ \frac{\partial}{\partial \mu} \ln p(X|\mu, \Sigma) = \sum_{n=1}^{N} \Sigma^{-1}(x_n - \mu) = 0 \]

- and solve to obtain

\[ \mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n. \]

- Similarly

\[ \Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{ML})(x_n - \mu_{ML})^T. \]
Mixtures of Gaussians (Bishop 2.3.9)

Old Faithful geyser:
The time between eruptions has a bimodal distribution, with the mean interval being either 65 or 91 minutes, and is dependent on the length of the prior eruption. Within a margin of error of ±10 minutes, Old Faithful will erupt either 65 minutes after an eruption lasting less than $2 \frac{1}{2}$ minutes, or 91 minutes after an eruption lasting more than $2 \frac{1}{2}$ minutes.
Mixtures of Gaussians (Bishop 2.3.9)

• Combine simple models into a complex model:

\[ p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) \]

\( \forall k : \pi_k \geq 0 \) \quad \sum_{k=1}^{K} \pi_k = 1
Mixtures of Gaussians (Bishop 2.3.9)
Mixtures of Gaussians (Bishop 2.3.9)

- Determining parameters $\pi$, $\mu$, and $\Sigma$ using maximum log likelihood

\[
\ln p(X|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \right\}
\]

Log of a sum; no closed form maximum.

- Solution: use standard, iterative, numeric optimization methods or the *expectation maximization* algorithm (Chapter 9).
Entropy 1.6

$$H[x] = - \sum_x p(x) \log_2 p(x)$$

Important quantity in
• coding theory
• statistical physics
• machine learning
Differential Entropy

Put bins of width $\Delta$ along the real line

$$\lim_{\Delta \to 0} \left\{ - \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) \, dx$$

For fixed $\sigma$, differential entropy maximized when

in which case

$$p(x) = \mathcal{N}(x | \mu, \sigma^2)$$

$$H[x] = \frac{1}{2} \left\{ 1 + \ln(2\pi \sigma^2) \right\}.$$
The Kullback-Leibler Divergence

P true distribution, q is approximating distribution

\[
\begin{align*}
KL(p\|q) &= -\int p(x) \ln q(x) \, dx - \left( -\int p(x) \ln p(x) \, dx \right) \\
&= -\int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} \, dx
\end{align*}
\]

\[
KL(p\|q) \simeq \frac{1}{N} \sum_{n=1}^{N} \left\{ -\ln q(x_n|\theta) + \ln p(x_n) \right\}
\]

\[
KL(p\|q) \geq 0 \quad KL(p\|q) \neq KL(q\|p)
\]