Recap: The Bayes Classifier

- The Bayes classifier, h_B, for the M-class case is:

  $h_B(\mathbf{X}) = \alpha_i \ \mathit{if}$

$$\sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X}) \leq \sum_{j=0}^{M-1} L(\alpha_k, C_j) q_j(\mathbf{X}), \ \forall k$$

Recap: The Bayes Classifier

- The Bayes classifier, h_B, for the M-class case is:

  $h_B(\mathbf{X}) = i \ if$

$$\sum_{j=0}^{M-1} L(i,j) q_j(\mathbf{X}) \leq \sum_{j=0}^{M-1} L(k,j) q_j(\mathbf{X}), \ \forall k$$

- For M-class case and 0–1 loss function, it is $h_B(\mathbf{X}) = i$ if

$$q_i(\mathbf{X}) \geq q_j(\mathbf{X}), \ \ \forall j$$

  or equivalently

$$p_i f_i(\mathbf{X}) \geq p_j f_j(\mathbf{X})$$

- This is optimal for minimizing Risk.

Recap

- We have seen many examples of computing Bayes classifier.
- For example, for Gaussian class conditional densities, Bayes classifier is a quadratic discriminant function. It is linear if all classes have the same covariance matrix.

Recap

- The Bayes classifier is optimal for the criterion of risk minimization.
- There can be other criteria.
- Minmax classifier has risk that is independent of priors.

Neyman-Pearson Criterion

- Bayes classifier minimizes risk.
- It minimizes some weighted sum of all errors.
- We may not explicitly want to trade one type of error with another
- Another criterion: minimize Type-II error under the constraint that Type-I error is below some given threshold.
- This is the Neyman-Pearson criterion.
- This could be useful in, e.g., biometric applications.

Neyman-Pearson Classifier

- For a given $\alpha \in (0,\ 1)$ as the bound on Type-I error, the Neyman-Pearson classifier, h_{NP}, is characterized by:

  1. $P[h_{NP}(\mathbf{X}) = 1 \,|\, \mathbf{X} \in \mathbf{C\text{-}0}] \leq \alpha$
  2. $P[h_{NP}(\mathbf{X}) = 0 \,|\, \mathbf{X} \in \mathbf{C\text{-}1}] \leq [P[h(\mathbf{X}) = 0 \,|\, \mathbf{X} \in \mathbf{C\text{-}1}]$
     for all h such that $P[h(\mathbf{X}) = 1 \,|\, \mathbf{X} \in \mathbf{C\text{-}0}] \leq \alpha$

- Type-I error: Wrongly classifying a Class-0 pattern.
- The Type-I error of NP classifier is bounded above by α.
- Among all classifiers that satisfy this bound on Type-I error, NP classifier has the least Type-II error.
- The Neyman Person classifier can also be expressed as a threshold on the likelihood ratio.

Neyman-Person Classifier

- Given the bound on Type-I error, α, the Neyman-Person Classifier can be shown to be

$$\begin{aligned} h_{NP}(\mathbf{X}) &= 1 \text{ if } \frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > K \\ &= 0 \text{ Otherwise} \end{aligned}$$

where K is such that

$$P\left[\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} \leq K \mid \mathbf{X} \in \mathbf{C\text{-}0}\right] = 1 - \alpha$$

(We assume $P\{\mathbf{X} : f_1(\mathbf{X}) = Kf_0(\mathbf{X})\} = 0$ for simplicity)

- We now prove that this satisfies the NP Criterion.
- The threshold K for the NP classifier is chosen so that

$$P\left[\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} \leq K \mid \mathbf{X} \in \mathbf{C\text{-}0}\right] = 1 - \alpha$$

- Hence, by construction, we have

$$\begin{aligned} P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}0}] &= P\left[\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > K \mid \mathbf{X} \in \mathbf{C\text{-}0}\right] \\ &= \alpha \end{aligned}$$

- So, we need to show that its Type-II error is less than that for any other classifier satisfying the constraint on Type-I error.

- Let h be any classifier such that

$$P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \textbf{C-0}] \leq \alpha$$

- To complete the proof we have to show that

$$P[h_{NP}(\mathbf{X}) = 0 \mid \mathbf{X} \in \textbf{C-1}] \leq P[h(\mathbf{X}) = 0 \mid \mathbf{X} \in \textbf{C-1}]$$

  Or, equivalently

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \textbf{C-1}] \geq [P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \textbf{C-1}]$$

- Consider the Integral

$$\begin{aligned} I &= \int_{\Re^n} (h_{NP}(\mathbf{x}) - h(\mathbf{x}))\,(f_1(\mathbf{x}) - Kf_0(\mathbf{x}))\,d\mathbf{x} \\ &= \int_{f_1 > Kf_0} (h_{NP}(\mathbf{x}) - h(\mathbf{x}))\,(f_1(\mathbf{x}) - Kf_0(\mathbf{x}))\,d\mathbf{x} + \\ &\qquad \int_{f_1 \leq Kf_0} (h_{NP}(\mathbf{x}) - h(\mathbf{x}))\,(f_1(\mathbf{x}) - Kf_0(\mathbf{x}))\,d\mathbf{x} \end{aligned}$$

- We first show that this integral is always non-negative.

- When $f_1(\mathbf{x}) > Kf_0(\mathbf{x})$, we have $h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 1 - h(\mathbf{x}) \geq 0$ which implies

$$(h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - Kf_0(\mathbf{x})) \geq 0$$

- Similarly, when $f_1(\mathbf{x}) < Kf_0(\mathbf{x})$, we have $h_{NP}(\mathbf{x}) - h(\mathbf{x}) = 0 - h(\mathbf{x}) \leq 0$ which implies

$$(h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - Kf_0(\mathbf{x})) \geq 0$$

- This shows that $I \geq 0$. That is,

$$I = \int_{\Re^n} (h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - Kf_0(\mathbf{x}))\, d\mathbf{x} \geq 0$$

- Thus, we have

$$\int_{\Re^n} (h_{NP}(\mathbf{x}) - h(\mathbf{x}))(f_1(\mathbf{x}) - Kf_0(\mathbf{x}))\ d\mathbf{x} \geq 0$$

- This implies

$$\int\ h_{NP}(\mathbf{x})f_1(\mathbf{x})\ d\mathbf{x} - \int\ h(\mathbf{x})f_1(\mathbf{x})\ d\mathbf{x} \geq$$

$$K\left[\int\ h_{NP}(\mathbf{x})f_0(\mathbf{x})\ d\mathbf{x} - \int\ h(\mathbf{x})f_0(\mathbf{x})\ d\mathbf{x}\right]$$

Since h_{NP} and h take values in $\{0,\ 1\}$,

$$\int_{\Re^n} h_{NP}(\mathbf{x}) f_1(\mathbf{x}) d\mathbf{x} = P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}]$$

and

$$\int_{\Re^n} h(\mathbf{x}) f_1(\mathbf{x}) d\mathbf{x} = P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}]$$

Similarly for the integrals involving f_0.

- We have shown

$$\int h_{NP}(\mathbf{x}) f_1(\mathbf{x})\, d\mathbf{x} - \int h(\mathbf{x}) f_1(\mathbf{x})\, d\mathbf{x} \geq$$

$$K \left[\int h_{NP}(\mathbf{x}) f_0(\mathbf{x})\, d\mathbf{x} - \int h(\mathbf{x}) f_0(\mathbf{x})\, d\mathbf{x} \right]$$

- Hence we have

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}] \geq$$

$$K \left[P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}0}\ ] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}0}] \right]$$

- Hence we have

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}] \geq$$

$$K\left[P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}0}\ ] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}0}]\right]$$

- But for all h under consideration, the RHS above is non-negative.
  Hence

$$P[h_{NP}(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}] - P[h(\mathbf{X}) = 1 \mid \mathbf{X} \in \mathbf{C\text{-}1}] \geq 0$$

- This completes the proof.

- Neymann-Pearson classifier also needs knowledge of class conditional densities.
- Like Bayes classifier, it also is based on the ratio $\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})}$.
- In Bayes classifier we say class-1 if $\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > \frac{p_0}{p_1} \frac{L(0,1)}{L(1,0)}$.
- In NP, this threshold, K, is set based on the allowed Type-I error.

Example of NP classifier

- Take $X \in \Re$ and class conditional densities normal with equal variance. Let $\mu_0 < \mu_1$.
- Now the NP classifier is: If $X > \tau$ then class-1 where τ is simply determined by Type-I error bound.
- We can show that τ is determined by $\int_{\tau}^{\infty} f_0(x)\, dx = \alpha$.
- SHOW IT!! – Home work.

- Bayes classifier minimizes risk which is a kind of weighted error.
- NP classifier is another way of deciding how to trade one type of error with another.
- For a 2-class problem, one can always trade false positive rate against false negative rate.
- Receiver operating characteristic (ROC) is one way of effecting this.

Probability of error Vs Threshold

- Consider a one dimensional feature space, 2-class problem with a classifier, $h(X) = 0$ if $X < \tau$.
- Consider equal priors, Gaussian class conditional densities with equal variance, $0 - 1$ loss.

  Now let us write the probability of error as a function of τ.

$$
\begin{aligned}
P[\text{error}] &= 0.5 \int_{-\infty}^{\tau} f_1(X)\, dX \;+\; 0.5 \int_{\tau}^{\infty} f_0(X)\, dX \\
&= 0.5\, \Phi\left(\frac{\tau - \mu_1}{\sigma}\right) \;+\; (1 \;-\; \Phi\left(\frac{\tau - \mu_0}{\sigma}\right))
\end{aligned}
$$

- As we vary τ we trade one kind of error with another. In Bayes classifier, the loss function determines the ‘exchange rate’.

- The receiver operating characteristic (ROC) curve is one way to conveniently visualize and exploit this trade off.
- For a two class classifier there are four possible outcomes of a classifcation decison – two are correct decisions and two are errors.
- Let e_i denote probability of wrongly assigning class i, $i = 0, 1$.

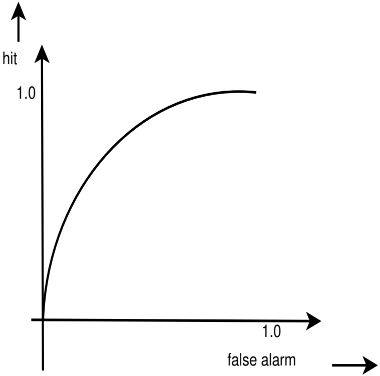
ROC curve

Then we have

$$\begin{aligned}
e_0 &= P[X \leq \tau \mid X \in \mathbf{c_1}] && \text{(a miss / false negative)} \\
e_1 &= P[X > \tau \mid X \in \mathbf{c_0}] && \text{(false alarm / false positive)} \\
1 - e_0 &= P[X > \tau \mid X \in \mathbf{c_1}] && \text{(true positive / hit)} \\
1 - e_1 &= P[X \leq \tau \mid X \in \mathbf{c_0}] && \text{(true negative)}
\end{aligned}$$

- For fixed class conditional densities, if we vary τ the point $(e_1,\ 1 - e_0)$ moves on a smooth curve in $\Re^2$.
- This is traditionally called the ROC curve. (Choice of coordinates is arbitrary)

Example ROC curve



- For any fixed τ we can estimate e_0 and e_1 from training data.
- Hence, varying τ we can find ROC and decide which may be the best operating point.
- This can be done for any threshold based classifier irrespective of class conditional densities.
- Estimates of e_0 and e_1 from data is often called false negative rate (FNR) and false positive rate (FPR). The $(1 - e_0)$ is referred to as the true positive rate (TPR).

- When the class conditional densities are Gaussian with equal variance, we use this procedure to estimate Bayes error also.
- When class conditional densities are normal with equal variances,

$$e_1 = \int_{\tau}^{-\infty} f_0(x)\, dx = 1 - \Phi\left(\frac{\tau - \mu_0}{\sigma}\right)$$

$$e_0 = \int_{-\infty}^{\tau} f_1(x)\, dx = \Phi\left(\frac{\tau - \mu_1}{\sigma}\right)$$

- From these error integrals we get

$$\begin{aligned}\frac{\tau - \mu_0}{\sigma} &= \Phi^{-1}(1 - e_1) = a, \text{ say} \\ \frac{\tau - \mu_1}{\sigma} &= \Phi^{-1}(1 - (1 - e_0)) = b, \text{ say}\end{aligned}$$

- Then, $|a - b| = \frac{|\mu_1 - \mu_0|}{\sigma} = d$, the discriminability. (This does not depend on τ).
- Knowing $e_1, (1 - e_0)$, we can get d and hence the Bayes error.
- For our given τ we can also get the actuall error probability. We can tweak τ to match the Bayes error.

- We can in general use the ROC curve in multidimensional cases also.
- Consider

$$h(\mathbf{X}) = \text{sgn}(\mathbf{W}^T\mathbf{X} + w_0).$$

  can use ROC to fix w_0 after learning $\mathbf{W}$.

- ROC is a plot of TPR versus FPR for various values of the threshold.
- The point (0, 1) represents the perfect classification.
- The line joining $(0, 0)$ to $(1, 1)$ would represent a random classifier.
- Shape of ROC characterizes the ‘difficulty’ of the problem.
- Area under ROC curve is used as a good indicator of the quality of classifier (irrespective of threshold)

Implementing Bayes Classifier

- We need class conditional densities and prior probabilities.
- Prior probabilities can be estimated as fraction of examples from each class.
- Since examples are *iid* and the class labels of examples are known, we have some iid samples from each class conditional distribution.
- The problem:

  Given $\{x_1, x_2, \cdots, x_m\}$ *drawn* iid *according to some distribution, estimate the probability distribution / density.*

Estimating densities

- Two main approaches: Parametric and non-parametric.
- Parametric: We assume we have *iid* realizations of a random variable X whose distribution is known except for values of a parameter vector. We estimate the parameters of the density using the samples available.
- In non-parametric approach we do not assume form of density. It is often modelled as a convex combination of some densities using the samples.

Estimating parameters of a density

- Denote the density by $f(x \mid \theta)$ where θ is a parameter vector.
- For example, let $\theta = (\theta_1,\ \theta_2)$ and

  $$f(x \mid \theta) = \frac{1}{2\pi\sqrt{\theta_2}} \exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right)$$

  $f(x|\theta)$ is normal with mean and variance constituting the parameter vector.
- Now estimation of density is same as estimation of a parameter vector.

Notation

- Let X denote a random variable with density $f(x \mid \theta)$. (Use same notation even when X is a random vector)
- A (*iid*) sample of size n consists of n *iid* realizations of X.
- $\mathbf{x} = (x_1, \cdots, x_n)^T$ – the sample or the data. We sometimes use $\mathcal{D}$ to denote the data.
- It can be thought of as a realization of $(X_1, \cdots, X_n)^T$ where X_i are *iid* with density $f(x \mid \theta)$.

- A *statistic* is a function of data, e.g., $g(x_1, \cdots, x_n)$.
- An estimator is such a statistic. $\hat{\theta}(x_1, \cdots, x_n)$.
- When we need to remember the sample size, we write $\hat{\theta}_n$
- For example,

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$$

the well-known sample mean.

- We need ‘good’ estimators.
- We need some criteria for ‘goodness’. Also, methods to obtain such estimators.
- In this course, we will consider two methods: Maximum likelihood and Bayesian estimators.
- To begin with a simple introduction to some general issues in estimation.

- An estimator, $\hat{\theta}$ of a parameter (vector) θ is said to be **unbiased** if $E[\hat{\theta}] = \theta$.
- The $\hat{\theta}$ is a function of data. Hence the expectation is with respect to the joint density of $(X_1, \cdots X_n)$.
- Since $X_i \sim f(x \mid \theta)$, the expectation above needs value of θ.
- So, to be more precise, unbiasedness requires

  $$E_\theta[\hat{\theta}] = \theta$$

  where E_θ denotes expectation with respect to the joint density of $(X_1, \cdots X_n)$ with X_i independent and $X_i \sim f(x \mid \theta)$.

- Let $\hat{\theta}_n = (1/n) \sum_i x_i$
- Then, for every n, $E_\theta[\hat{\theta}_n] = \theta$ because $E_\theta X_i = \theta$, $\forall \theta$.
- Sample mean is an unbiased estimator of actual mean.
- Let $\hat{\theta}'(x_1, \cdots, x_n) = 0.5(x_1 + x_2)$.
- This is also an unbiased estimator. So is $\hat{\theta}'' = x_1$.
- Unbiasedness alone is not enough

- The mean square error of an estimator is defined by

$$\mathsf{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2]$$

- Result:

  $$\mathrm{MSE}_\theta(\hat{\theta}) = V_\theta(\hat{\theta}) + [B_\theta(\hat{\theta})]^2$$

  where $V_\theta(\hat{\theta})$ is the **variance** given by

  $$V_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - E_\theta[\hat{\theta}])^2]$$

  and $B_\theta(\hat{\theta})$ is the **bias** given by

  $$B_\theta(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$$

- For unbiased estimators, the bias is zero.
- For unbiased estimators the variance is the mean square error.

- Proof:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta}-\theta)^2] \\
&= E[\{(\hat{\theta}-E[\hat{\theta}]) + (E[\hat{\theta}]-\theta)\}^2] \\
&= E[(\hat{\theta}-E[\hat{\theta}])^2] + (E[\hat{\theta}]-\theta)^2 + \\
&\qquad 2E\left[(\hat{\theta}-E[\hat{\theta}])(E[\hat{\theta}]-\theta)\right] \\
&= V(\hat{\theta}) + [B(\hat{\theta})]^2 + 2(E[\hat{\theta}]-\theta)E[(\hat{\theta}-E[\hat{\theta}])] \\
&= V(\hat{\theta}) + [B(\hat{\theta})]^2
\end{aligned}$$

- For unbiased estimators, low variance implies low MSE.
- Earlier example: When $\hat{\theta}_n = (1/n)\sum_i x_i$, the sample mean,

$$V_\theta(\hat{\theta}_n) = (1/n)^2 \sum_i \mathrm{Var}(x_i) = \frac{\sigma^2}{n}$$

- For $\hat{\theta}'_n = 0.5(x_1 + x_2)$,

$$V_\theta(\hat{\theta}'_n) = \frac{\sigma^2}{2}$$

- Hence $\hat{\theta}$ is better than $\hat{\theta}'$
  (under the criterion of MSE)

- So, unbiased estimators with low variance are good.
- For a given family of density functions, $\hat{\theta}$ is said to be **uniformly minimum variance unbiased estimator (UMVUE)** if
    1. $\hat{\theta}$ is unbiased, and
    2. $\mathsf{MSE}_\theta(\hat{\theta}_n) \leq \mathsf{MSE}_\theta(\hat{\theta}'_n)\ \forall n, \theta$,
       and forall $\hat{\theta}'$ that are unbiased estimators for θ.
- If we can get an UMVUE, then it is the 'best' estimator.
- In many cases, it is difficult to get UMVUE.

- So far, we are looking at figures of merit of estimators at (all) fixed sample sizes.
- We can also think of asymptotic properties.
- An estimator $\hat{\theta}$ is said to be **consistent** for θ if

$$\hat{\theta}_n \xrightarrow{P} \theta \ \ \forall\theta$$

- For example, the sample mean is a consistent estimator of population mean (expectation of the random variable) (Law of large numbers)

- A consistent estimator need not be unbiased.
- Let θ be the mean and let

$$\hat{\theta}_n = \frac{1}{n+1} \sum_{i=1}^{n} x_i$$

- This is not an unbiased estimator. But it is easy to show that $E[(\hat{\theta}_n - \theta)^2] \to 0$ as $n \to \infty$.

- Maximum Likelihood (ML) estimation is a general procedure for obtaining consistent estimators.
- It is a parametric method.
- We estimate parameters of a density based on *iid* samples.
- For most densities, ML estimates are consistent.

Maximum likelihood estimation

- Likelihood function is defined by

$$L(\theta, \mathbf{x}) = \prod_{j=1}^{n} f(x_j|\theta)$$

- We essentially look at the likelihood function as a function of θ with the x_j being known values (as given by data).
- To emphasize this we write it as $L(\theta \mid \mathbf{x})$ or $L(\theta \mid \mathcal{D})$.

Maximum likelihood estimation contd..

- The maximum likelihood (ML) estimate of θ is the value that (globally) maximizes the likelihood function.
- θ^* is the MLE for θ if

$$L(\theta^* \mid \mathbf{x}) \geq L(\theta \mid \mathbf{x}) \;\; \forall \theta$$

- Finding MLE is an optimization problem.

- For convenience in optimization we often take the log likelihood given by

$$l(\theta \mid \mathbf{x}) = \log L(\theta \mid \mathbf{x}) = \sum_{j=1}^{n} \log f(x_j|\theta)$$

- Now the ML estimate would be maximizer of the log likelihood.
- For many densities we can analytically solve for the maximizer.
- In general we can use numerical optimization techniques.

Example

- Consider one dimensional case.
  Let $f(x|\theta) \sim \mathcal{N}(\mu, \sigma^2)$ with $\theta_1 = \mu$ and $\theta_2 = \sigma$.

$$f(x|\theta) = \frac{1}{\theta_2\sqrt{2\pi}} exp\left(-\frac{(x-\theta_1)^2}{2\theta_2^2}\right)$$

- Now log likelihood would be

$$\begin{aligned} l(\theta|\mathbf{x}) &= \sum_{j=1}^{n} \log f(x_j|\theta) \\ &= \sum_{j=1}^{n} \left[-\log(\theta_2) - 0.5\log(2\pi) - \frac{(x_j-\theta_1)^2}{2\theta_2^2}\right] \\ &= -n\log(\theta_2) - 0.5n\log(2\pi) - \sum_{j=1}^{n} \frac{(x_j-\theta_1)^2}{2\theta_2^2} \end{aligned}$$

- The log likelihood is given by

$$l(\theta|\mathbf{x}) = -n\log(\theta_2) - 0.5n\log(2\pi) - \sum_{j=1}^{n} \frac{(x_j - \theta_1)^2}{2\theta_2^2}$$

- To maximize log likelihood we equate the partial derivatives to zero.

$$\begin{aligned} \frac{\partial l}{\partial \theta_1} &= \frac{1}{\theta_2^2}\sum_{j=1}^{n}(x_j - \theta_1) = 0 \\ \frac{\partial l}{\partial \theta_2} &= -\frac{n}{\theta_2} + \frac{1}{\theta_2^3}\sum_{j=1}^{n}(x_j - \theta_1)^2 = 0 \end{aligned}$$

- Solving these, we get

$$\frac{\partial l}{\partial \theta_1} = \frac{1}{\theta_2^2}\sum_{j=1}^{n}(x_j - \theta_1) = 0 \quad \Rightarrow \quad \hat{\theta}_1 = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\frac{\partial l}{\partial \theta_2} = -\frac{n}{\theta_2}+\frac{1}{\theta_2^3}\sum_{j=1}^{n}(x_j-\theta_1)^2 = 0 \quad \Rightarrow \quad \frac{1}{\theta_2^3}\sum_{j=1}^{n}(x_j-\theta_1)^2 = \frac{n}{\theta_2}$$

$$\Rightarrow \quad \hat{\theta}_2^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\theta}_1)^2$$

- These are the ML estimates of mean and variance of a normal density
- ML estimate of variance is **not** unbiased.

Example: discrete case

- Let X have Bernoulli distribution. That is, X takes values 0 and 1 with probability $(1-p)$ and p respectively.

$$f(x|p) = p^x(1-p)^{1-x},\ x \in \{0,1\}$$

- The mass function has only one parameter, namely, p.
- The likelihood function is

$$L(p|\mathbf{x}) = \prod_{j=1}^{n} f(x_j|p) = \prod_{j=1}^{n} p^{x_j}(1-p)^{1-x_j} = p^{n\bar{x}}(1-p)^{n-n\bar{x}}$$

where $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ is the sample mean.

- The likelihood function is given by

$$L(p|\mathbf{x}) = p^{n\bar{x}}(1-p)^{n-n\bar{x}}$$

- The loglikelihood is given by

$$l(p|\mathbf{x}) = n\bar{x}\log p + n(1-\bar{x})\log(1-p)$$

- Differentiating with respect to p and equating to zero

$$\frac{n\bar{x}}{p} = \frac{n(1-\bar{x})}{1-p}$$

which implies

$$p = \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

- The sample mean is the ML estimate of the parameter p of a Bernoulli random variable.

Another Example

- Consider the multidimensional Gaussian density

$$f(x \mid \theta) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

  where $x \in \Re^d$ and $\theta = (\mu, \Sigma)$ are the parameters.

- For a random vector X having the above joint density, $\mu \in \Re^d$ is the mean vector (i.e., $EX = \mu$) and the $d \times d$ matrix Σ is the covariance matrix defined by

$$\Sigma = E(X-\mu)(X-\mu)^T$$

The log likelihood function is given by

$$
\begin{aligned}
l(\theta \mid \mathcal{D}) &= \sum_{i=1}^{n} \ln\left(f(x_i|\theta)\right) \\
&= \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \ \exp\left(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)\right) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2}\ln((2\pi)^d|\Sigma|) - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)
\end{aligned}
$$

where $\theta = (\mu, \Sigma)$ constitute the parameters to be estimated.

- To find the ML estimates, we have to equate the partial derivatives of l (with respect to the parameters) to zero and solve.

$$l(\mu, \Sigma|\mathcal{D}) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln((2\pi)^d |\Sigma|) - \frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right)$$

- Now, $\frac{\partial l}{\partial \mu} = 0$ gives us

$$\sum_{i=1}^{n} \Sigma^{-1}(x_i - \mu) = 0$$

- This gives us the ML estimate for μ as

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Thus, even in the multidimensional case, the ML estimate for mean is the sample mean.

- Finding the partial derivative with respect to Σ is algebraically involved.
- We can show that the ML estimate for Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

- Again, the final ML estimate is intuitively obvious. (Recall that $\Sigma = E(X - \mu)(X - \mu)^T$).

ML estimate for Σ

- The log likelihood is

$$l(\Sigma, \mu|\mathcal{D}) = \sum_{i=1}^{n} \left(-\frac{1}{2} \ln((2\pi)^d |\Sigma|) - \frac{1}{2}(x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right)$$

- We use $\Lambda = \Sigma^{-1}$ as the parameter. It is called precision matrix.
- Now we can rewrite log likelihood as (using $z_i = (x_i - \mu)$)

$$l(\Lambda, \mu \mid \mathcal{D}) = -\frac{nd}{2} \ln(2\pi) + \frac{n}{2} \ln(|\Lambda|) - \frac{1}{2} \sum_{i=1}^{n} z_i^T \Lambda z_i$$

- So, we need derivatives of the form

$$\frac{\partial z^T A z}{\partial A} \quad \text{and} \quad \frac{\partial \ln(|A|)}{\partial A}$$

- Some useful matrix identities:

$$\begin{aligned}
\mathrm{Tr}(AB) &= \mathrm{Tr}(BA) \\
\mathrm{Tr}(ABC) &= \mathrm{Tr}(CAB) = \mathrm{Tr}(BCA) \\
\frac{\partial \mathrm{Tr}(AB)}{\partial A} &= B^T \\
\frac{\partial \mathrm{Tr}(A^T B)}{\partial A} &= B
\end{aligned}$$

where $\mathrm{Tr}(A)$ is the trace of the (square) matrix A.

- Using these we have

$$z^T Az = \mathrm{Tr}(z^T Az) = \mathrm{Tr}(zz^T A) = \mathrm{Tr}(Azz^T)$$

- Hence we have

$$\frac{\partial z^T Az}{\partial A} = \frac{\partial \mathrm{Tr}(Azz^T)}{\partial A} = (zz^T)^T = zz^T$$

- We could derive this from first principles also.

$$z^T A z = \sum_{k,l} A_{kl} z_k z_l$$

- Hence

$$\frac{\partial z^T A z}{\partial A_{ij}} = z_i z_j$$

- Thus we have

$$\frac{\partial z^T A z}{\partial A} = z z^T$$

- The other derivative we need is $\frac{\partial \ln(|A|)}{\partial A}$.
- For this we use the following identity

$$\frac{\partial \ln(|A|)}{\partial x} = \mathrm{Tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)$$

- Proving this identity is a bit involved.

- Using

$$\frac{\partial \ln(|A|)}{\partial x} = \mathrm{Tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)$$

  we get

$$\frac{\partial \ln(|A|)}{\partial A_{ij}} = \mathrm{Tr}\left(A^{-1}\frac{\partial A}{\partial A_{ij}}\right)$$

- $\frac{\partial A}{\partial A_{ij}}$ is a matrix with '1' in position (i, j) and zeroes elsewhere.

- Thus we have

$$\begin{aligned}\left(A^{-1}\frac{\partial A}{\partial A_{ij}}\right)_{kl} &= \sum_{s} A^{-1}_{ks}\left(\frac{\partial A}{\partial A_{ij}}\right)_{sl} \\ &= \begin{cases} 0 & \text{if } \ l \neq j \\ A^{-1}_{ki} & \text{if } \ l = j \end{cases}\end{aligned}$$

- Hence we get

$$\mathrm{Tr}\left(A^{-1}\frac{\partial A}{\partial A_{ij}}\right) = \left(A^{-1}\frac{\partial A}{\partial A_{ij}}\right)_{jj} = A^{-1}_{ji}$$

- Thus we get

$$\frac{\partial \ln(|A|)}{\partial A_{ij}} = \mathrm{Tr}\left(A^{-1}\frac{\partial A}{\partial A_{ij}}\right) = A_{ji}^{-1}$$

- Hence

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

ML estimate of Σ

- The log likelihood is (with $z_i = (x_i - \mu)$)

$$l(\Lambda, \mu \mid \mathcal{D}) = -\frac{nd}{2}\ln(2\pi) + \frac{n}{2}\ln(|\Lambda|) - \frac{1}{2}\sum_{i=1}^{n} z_i^T \Lambda z_i$$

- Hence we have

$$\frac{\partial l(\Lambda, \mu \mid \mathcal{D})}{\partial \Lambda} = \frac{n}{2}\Lambda^{-1} - \frac{1}{2}\sum_{i=1}^{n} z_i z_i^T$$

- Equating the partial derivative to zero we have

$$\frac{\partial l(\Lambda, \mu \mid \mathcal{D})}{\partial \Lambda} = \frac{n}{2}\Lambda^{-1} - \frac{1}{2}\sum_{i=1}^{n} z_i z_i^T = 0$$

- Solving this, we get

$$\Sigma = \Lambda^{-1} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^T$$

- As we saw earlier this is intuitively obvious.