EE126: Probability and Random Processes

Lecture 25: Classical Statistical Inference

Abhay Parekh

UC Berkeley

April 26, 2011

Logistics

- Last class with new material
- Bspace will be up to date by Thursday
- Thursday lecture is a review for the final.
- GSI's will have review sessions next week
- Final is on May 13

Two ways to treat unknowns

- The goal of the inference is to come up with the best estimate of the VALUE of the unknown: Example: The bias of the coin is 0.75.
- The goal of the inference is to come up with the best estimate of the DISTRIBUTION of the unknown: Example: The bias of the coin is uniformly distributed between $[0.75, 0.85]$.

The first approach follows CLASSICAL statistics and the second approach follows BAYESIAN statistics.

$E[\Theta|X]$ is the estimator with the lowest MSE.

The estimation error is $\bar{\Theta} = \hat{\theta} - \theta$. Let $\hat{\theta} = E[\Theta|X]$:

1. $E[\bar{\theta}] = 0$ Error is unbiased
2. $\mathrm{cov}(\hat{\theta}, \bar{\theta}) = 0$: error is uncorrelated with the estimate.
3. $var(\Theta) = var(\hat{\theta}) + \mathrm{var}(\bar{\theta})$

Estimating Multiple Parameters and Multiple Observations

We need to estimate Θ_i, $i = 1, 2, ..., k$. What to do?
Minimize the sum of the individual MSE's:

$$E[\Theta_1 - \hat{\theta}_1] + \cdots + E[\Theta_k - \hat{\theta}_k]$$

Solve k decoupled MSE problems: $\hat{\theta}_i = E[\Theta_i|X]$.
Multiple Observations: Computation gets tough...

Bayes LLSE with one observation

- Given an observation X:

$$\hat{\Theta} = E[\Theta] + \frac{cov(X,\Theta)}{\sigma_X^2}(X - E[X]) = E[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

  where

$$\rho_{\Theta,X} = \frac{cov(\Theta,X)}{\sigma_X\sigma_\Theta}$$

- The MSE is

$$(1-\rho^2)\sigma_\Theta^2.$$

Properties of Estimators

Given observations $X = (X_1, ..., X_n)$ we use the estimation rule $\hat{\Theta} = g(X)$. Since we do not want to assume a prior for Θ, $g(X)$ must work "well" for **all** possible values of Θ.

1. Bias (Mean estimation error): $b_\theta(\hat{\Theta}_n) = E[\hat{\theta}_n] - \theta$.
   - If bias is **zero for all** θ, the estimator is **unbiased.**
   - If $\lim_{n\to\infty} b_\theta(\hat{\Theta}_n) = 0$ for all θ, the estimator is **asymptotically unbiased**.
2. If $\hat{\Theta}_n$ converges to the true value of θ in probability for all θ, the estimator is **consistent**.

Example: Estimating the Mean of a Random Variable, X

Given iid observations $X_1, ..., X_n$, Estimation error: $M_n - X$.

$$E[M_n] = E[\sum_i X_i/n] = \sum_i E[X_i]/n = E[X]$$

Thus M_n is an unbiased estimator of the mean.

$$E[(M_n - E[X])^2] = E[(M_n - E[M_n])^2] = var(M_n) = \frac{var(X)}{n}$$

Is M_n the estimator with the smallest variance? No. Pick $\hat{\theta} = 0$!
But MSE $= \theta^2$.

$$\text{MSE} = b_\theta^2(\hat{\theta}) + var(\theta_n)$$

Example: Gaussian Observations

$X_1, ..., X_n$ are iid $N(\theta, v)$ where θ is unknown but v is known. We showed that if the prior on θ is $N(\theta, v)$, the LMS estimator:

$$\hat{\theta}_L = \frac{x_0 + X_1 + X_2 + \cdots + X_n}{n+1}$$

But if x_0 is zero (i.e. the prior mean is zero):

$$\hat{\theta}_L = \frac{X_1 + X_2 + \cdots + X_n}{n+1}$$

In this case, $E[\hat{\theta}_L] = \frac{n\theta}{n+1}$ and the estimator is biased. But it is asymptotically unbiased.

$$var(\hat{\Theta}_n) = \frac{vn}{(n+1)^2}$$

Thus, the bias is > 0 but the variance is slightly smaller than that of the unbiased estimator M_n.

Maximum Likelihood Estimation

ML Rule:

$$\hat{\theta} = \text{argmax}_\theta p_X(x_1, ..., x_n, \theta) \quad (\Theta \text{ discrete})$$

$$\hat{\theta} = \text{argmax}_\theta f_X(x_1, ..., x_n, \theta) \quad (\Theta \text{ continuous})$$

$\hat{\theta}$ is the most likely value of θ given the observations.
Two ways to look at this:

1. We are not using a prior distribution since we compute $\hat{\theta}$ without using it.
2. We are assuming that the prior is uniform (flat). Consider the MAP rule:

$$\hat{\theta}_M = \text{argmax}_\theta p_{\Theta|X}(\theta|x) = \text{argmax}_\theta \frac{p_X(x_1, ..., x_n, \theta)}{1/n}$$

which is the Maximum Likelihood rule.

Estimating the Mean and Variance of a Gaussian

The observations $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$. We want the ML estimate for the mean and variance.
The likelihood function:

$$f_X(x) = (\frac{1}{\sqrt{2\pi v}})^n \prod_{i=1}^{n} e^{-(x_i-\mu)^2/2v} = (\frac{1}{\sqrt{2\pi v}})^n e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2v}$$

$$\log f_X(x) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log v - \sum_{i=1}^{n}(x_i-\mu)^2/2v$$

ML Rule: Find the μ and v that minimize

$$\frac{n}{2}\log v + \sum_{i=1}^{n}(x_i-\mu)^2/2v$$

Estimating the Mean and Variance of a Gaussian

Let m_n be the realized value of M_n. Then

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2 &= \sum_{i=1}^{n}(x_i-m_n+m_n-\mu)^2 \\ &= \sum_{i=1}^{n}(x_i-m_n)^2+(m_n-\mu)^2-2(x_i-m_n)(m_n-\mu)\end{aligned}$$

But $\sum_{i=1}^{n}(x_i-m_n)(m_n-\mu)=(m_n-\mu)\sum_{i=1}^{n}(x_i-m_n)=0$. So

$$\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2v}=\sum_{i=1}^{n}\frac{1}{2v}((x_i-m_n)^2+(m_n-\mu)^2)=\frac{ns_n^2}{2v}+\frac{n(m_n-\mu)^2}{2v}$$

where s_n^2 is the realized value of $(\sum_{i=1}^{n}(X_i-M_n)^2)/n$.
Taking partials wrt μ and v, setting equal to zero and solving we get that the ML estimates: $\theta_n=(m_n,s_n^2)$. Therefore the ML estimator is $\hat{\Theta}_n=(M_n,S_n^2)$.

Invariance of the ML Estimator

Suppose we want to estimate $h(\theta)$ a one-to-function of θ. Suppose it is increasing...Then for all α:

$$P(h(\theta) \leq \alpha) = P(\theta \leq h^{-1}(\alpha))$$

so the estimate of $h(\theta)$ is just $h(\hat{\theta})$.
For MAP we have to find the pdf or pmf for $h(\Theta)$ which can be quite different.

Uniform Random Variable

$x = x_1, ..., x_n$ are n observations of a random variable distributed on $[0, \theta]$. What is the ML for θ?
The likelihood function is $\frac{1}{\theta^n}$. Clearly, $\theta \geq \max\{x_1, ..., x_n\}$. Thus the likelihood fxn is maximized at the min value, i.e.

$$\hat{\theta}_n = \max\{x_1, ..., x_n\}.$$

As $n \to \infty$, $\hat{\theta}_n \to \theta$ in probability: $\hat{\theta}_n$ is **consistent.**
Is $\hat{\theta}_n$ biased?Find $E[\hat{\theta}_n]$:$F_{\hat{\Theta}_n}(x) = \frac{x^n}{\theta^n}$ if $x \in [0, \theta]$.So

$$f_{\hat{\Theta}_n}(x) = \frac{nx^{n-1}}{\theta^n}$$

$$E[\hat{\Theta}_n] = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{\theta^n} \frac{\theta_n^{n+1}}{n+1} = \theta \frac{n}{n+1}$$

So $\hat{\theta}$ is biased, but asymptotically unbiased.

Linear Regression

We want the best linear linear relationship between rvs X and Y.

$$y \approx \Theta_0 + \Theta_1 x$$

Given n observations $x_1, ..., x_n$ and $y_1, ..., y_n$ find $\hat{\theta}_0$ and $\hat{\theta}_1$ which minimize the sum of the squares of the residuals.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Diff wrt θ_1 and setting equal to zero:

$$2\sum_i (y_i - \theta_0 - \theta_1 x_i)(-x_i) = 0 \Rightarrow \sum_i x_i y_i - x_i\theta_0 = \theta_1 \sum_i x_i^2$$

Similarly for θ_0 and Simplifying:

$$\boxed{\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{\sum_{i=1}^{n}(x - \bar{x})^2}}$$

$$\boxed{\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

Justification of Linear Regression

1. $$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2}$$
$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$$

   And Linear LMS:

$$\hat{\Theta} = E[\Theta] + \frac{cov(X,\Theta)}{\sigma_X^2}(X - E[X])$$

   Thus, Linear Regression gives us the Bayes Linear LSE!
2. Suppose the x_i are not random variables and we observe:

$$Y_i = \Theta_0 + \Theta_1 x_i + W_i, \quad i = 1, 2, ..., n$$

   where $W_i \sim N(0, \sigma^2)$, iid. Then $Y_i \sim N(x_i, \sigma^2)$ are independent. ML rule finds the Θ_0 and Θ_1 that maximizes:

$$f_Y(y) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2\sigma^2}\}$$

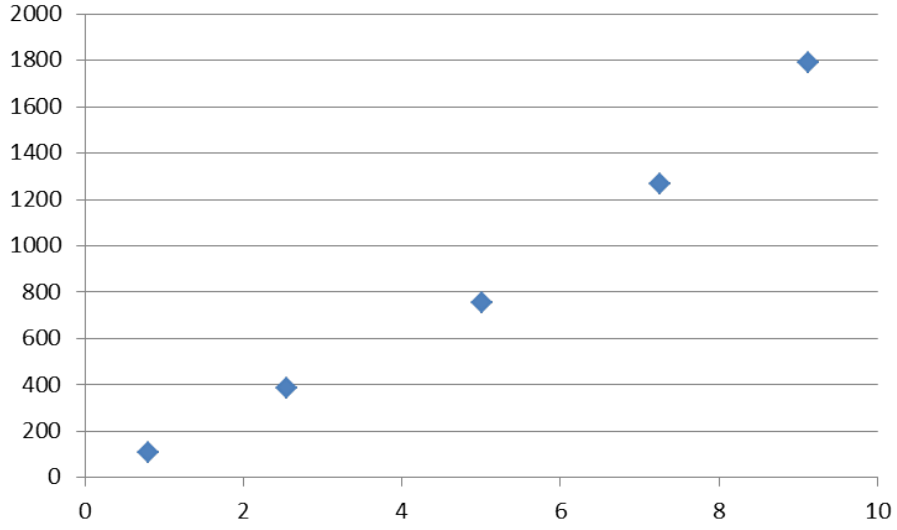
   But this is the same as minimizing the residuals!

Example: Which regression?

Alice has collected five data points $(x_i, y_i), i = 1, 2, 3, 4, 5$ for an important experiments.

x	0.798	2.546	5.005	7.261	9.131
y	106.141	387.19	753.986	1267.241	1789.137

She is trying to decide whether the relationship between X and Y is linear or quadratic by using a ML test. She assumes that her measurements are corrupted by iid noise with mean 0 and variance σ^2.



Approach

1. She finds the best fit for each model by minimizing the MSE. If linear:

$$Y_i = \theta_0 + \theta_1 X_i + W_i$$

If quadratic:

$$Y_i = \phi_0 + \phi X_i^2 + W_i$$

MSE: $\min_{\phi_0,\phi_1} \sum_{i=1}^{5}(y_i - \phi_0 + \phi_1 x_i^2)$Same form as linear wrt ϕ_0 and ϕ_1! So formulas are easy.

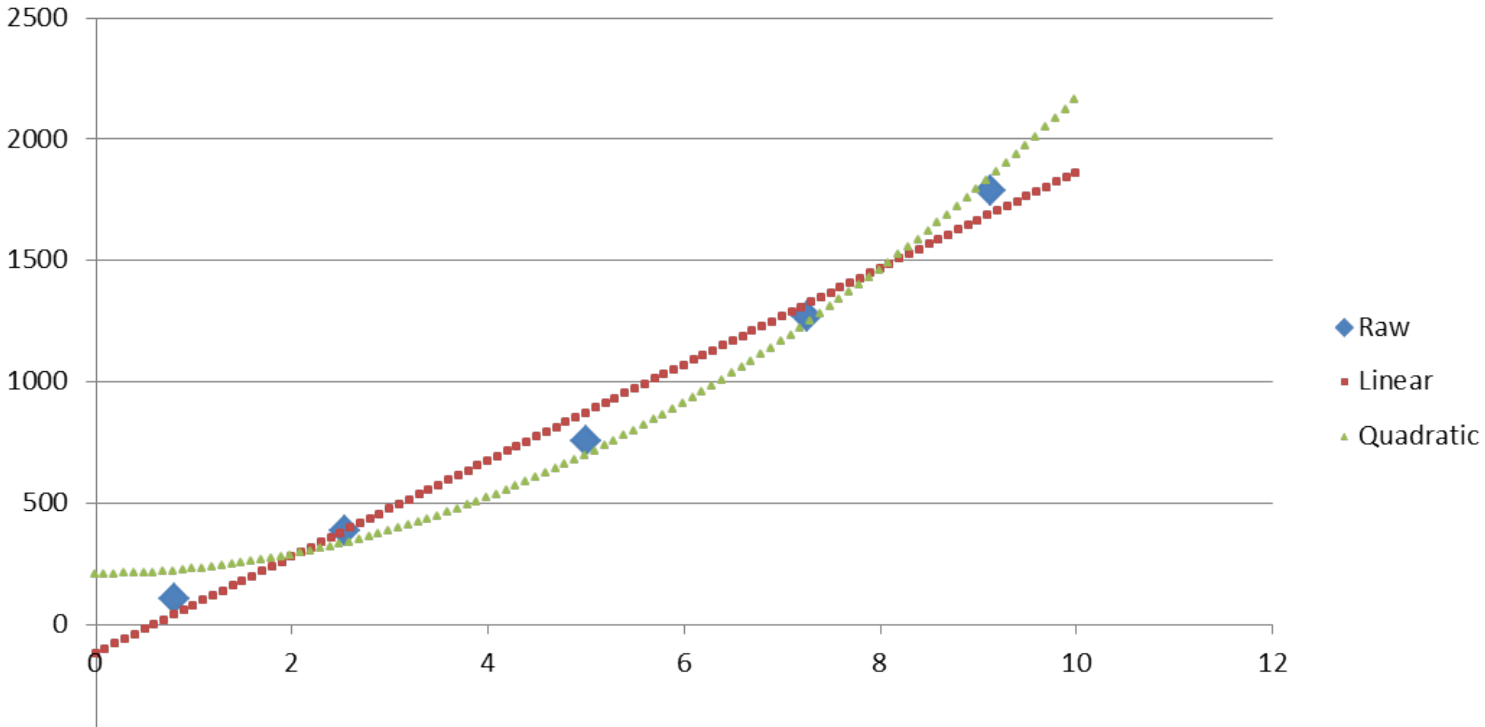
2. She compares the two hypotheses under a uniform prior.

Example: Which regression?

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{i=1}^{n}(x-\bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 x$$

$$\hat{\phi}_1 = \frac{\sum_{i=1}^{n}(x^2-\bar{x}^s)(y-\bar{y})}{\sum_{i=1}^{n}(x^2-\bar{x}^s)^2}, \quad \hat{\phi}_0 = \bar{y} - \hat{\phi}_1 \bar{x}^s$$

where $\bar{x}^s = \frac{1}{n}\sum_{i=1}^{5} x_i^2$.



ML Test

The likelihoods

$$f_{Y_1,\ldots,Y_5}(y_1,\ldots,y_5|\text{Linear}) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i-\hat{\theta}_0-\hat{\theta}_1 x_i)^2}{2\sigma^2}}$$

$$f_{Y_1,\ldots,Y_5}(y_1,\ldots,y_5|\text{Quadratic}) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i-\hat{\phi}_0-\hat{\phi}_1 x_i^2)^2}{2\sigma^2}}$$

By inspection, ML will pick linear iff

$$\boxed{\sum_i^5 (y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i)^2 \leq \sum_i^5 (y_i - \hat{\phi}_0 - \hat{\phi}_1 x_i^2)^2}$$

We find that for the data the LHS is $31,120.18$ and the RHS is $22,048.65$ so Alice decides that the relationship is **quadratic**.

Likelihood Ratio Test

Suppose there are two possible values: θ_0 and θ_1. Then MAP picks θ_1 iff $p_{\Theta|X)}(\theta_1|x) \geq p_{\Theta|X}(\theta_0|x)$, I.e.,

$$\frac{P(X = x|\theta = \theta_1)}{P(X = x|\theta = \theta_0} \geq \frac{P(\theta_0)}{P(\theta_1)}$$

$$L(x) \geq \frac{P(\theta_1)}{P(\theta_0)}$$

More general approach: Fix some η and pick θ_1 iff

$$L(x) \geq \eta$$

This is called a Likelihood Ratio Test.
If $\eta = 1$ we are using the Maximum Likelihood Rule.

Types of Error

MAP minimizes the overall prob of error. But there are two kinds in any hypothesis test:

- False Rejection (((α): θ_0 is the true value, but test picks θ_1
- False Acceptance (β): θ_1 is the true value but the test picks θ_0.

In a Likelihood Ratio Test: Increasing $\eta \Rightarrow$ Decreasing α and increasing β.

The Neyman Pearson Lemma

Suppose the goal is to find a test that minimizes β subject to the constraint that $\alpha < \alpha_0$, for some $\alpha \geq 0$. What kind of test is optimal?

Likelihood Ratio Tests are Powerful!

Suppose for a particular choice of η in the LRT it is true that

$$P(L(X) > \eta, \theta_0) = \alpha, \quad P(L(X) \leq \eta, \theta_1) = \beta.$$

Then for any other test that has rejection region R such that $P(\theta_0, X \in R) \leq \alpha$, it must be true that

$$P(\theta_1, X \notin R) \geq \beta.$$

Further if $P(\theta_0, P(X \in R)) < \alpha$ then

$$P(\theta_1, X \notin R) > \beta.$$

Proof of Neyman Pearson Lemma

Let $\frac{P(\theta_2)}{P(\theta_1)} = \eta$ so that $p_\Theta(\theta_0) = \eta p_\Theta(\theta_1) = 1 - p_\Theta(\theta_1)$.I.e.

$$p_\Theta(\theta_0) = \frac{\eta}{1+\eta}, \quad p_\Theta(\theta_1) = \frac{1}{1+\eta}$$

Now:

$$e_{MAP} = \frac{\eta}{1+\eta}\alpha + \frac{1}{1+\eta}\beta$$

We know that MAP maximizes the overall error. So for any other rule, i.e. any other region R:

$$e_{MAP} \leq \frac{\eta}{1+\eta}P(\theta_0, X \in R) + \frac{1}{1+\eta}P(\theta_1, X \notin R)$$

$$\frac{\eta}{1+\eta}(\alpha - P(\theta_0, X \in R) \leq \frac{1}{1+\eta}(P(\theta_1, X \notin R) - \beta)$$

Now if

- $P(\theta_0, P(X \in R) \leq \alpha \Rightarrow P(\theta_1, X \notin R) \geq \beta$
- $P(\theta_0, P(X \in R) < \alpha \Rightarrow P(\theta_1, X \notin R) > \beta$

Example: Predicting High Volume

The number of phone calls received by a ticket agency on any given day is a Poisson rv. On a normal day the rate is λ_0 but on popular days it is $\lambda_1 > \lambda_0$. Suppose they received k calls on a day. What is the LRT to determine if it is a normal day?

$$L(k) = \frac{\lambda_1^k e^{-\lambda_1}}{\lambda_0^k e^{-\lambda_0}} = (\frac{\lambda_1}{\lambda_0})^k e^{-(\lambda_1 - \lambda_0)}$$

$$L(k) \geq \eta \Rightarrow k(\log \lambda_1 - \log \lambda_0) - (\lambda_1 - \lambda_0) \geq \log \eta$$

So it is a popular day iff

$$k \geq \frac{\log \eta + \lambda_1 - \lambda_0}{\log \lambda_1 - \log \lambda_0}$$

Example: Bias of a Coin

The bias of a coin may be $\theta_0 = 0.5$ or $\theta_1 = 0.6$. For some large number of tosses, n, we obtain X_n heads. Devise an LRT to decide the bias.

$$L(k) = \frac{\binom{n}{k}0.6^k 0.4^{n-k}}{\binom{n}{k}0.5^k 0.5^{n-k}} = 1.2^k 0.8^{n-k} = 1.5^k 0.8^n$$

Since $L(k)$ is monotonically increasing with k, the test will accept the bias of 0.6 iff $X_n \geq k_n$.
Find k_n such that the prob of false rejection ≤ 0.05:
False rejection = Bias is 0.5 but test says Bias = 0.6.

$$P(H_0, X_n > k_n) \leq 0.05 \Rightarrow \sum_{i=k_n}^{n} \binom{n}{i} 0.5^i 0.5^{n-i} \leq 0.05$$

Now since n is large we can use the Laplace- Moivre Normal Approximation.

$$1 - \Phi(\frac{k_n - \frac{1}{2} - 0.5n}{\sqrt{n\ 0.5\ 0.5}}) \leq 0.05$$

Noting that $\Phi(1.644853) = 0.95$ we get

$$k_n = \frac{1}{2}(n+1) + 0.822427\sqrt{n}$$

Example: Bias of a Coin

What is the smallest value n for which both α and β and less than 0.05.

$$P(H_1, X_n > k_n) \leq 0.05 \Rightarrow \sum_{i=k_n}^{n} \binom{n}{i} 0.6^i 0.4^{n-i} \leq 0.05$$

Now since n is large we can use the Laplace- Moivre Normal Approximation.

$$1 - \Phi\left(\frac{k_n - \frac{1}{2} - 0.6n}{\sqrt{n\ 0.6\ 0.4}}\right) \leq 0.05$$

Substitute $k_n = \frac{1}{2}(n+1) + 0.822427\sqrt{n}$ and $\Phi(1.644853) = 0.95$ to get

$$n \geq 266$$

Minimum Expected Cost Hypothesis Testing

Let c_{ij} be the cost of the rule picking hypothesis i when hypothesis j is true.
Then given observation y, the expected cost of picking hypothesis i, is

$$E[i|y] = c_{i0}P(H_0|y) + c_{i1}P(H_1|y).$$

So given two hypotheses H_0 and H_1: Pick H_1 iff

$$c_{10}P(H_0|y) + c_{11}P(H_1|y) \leq c_{00}P(H_0|y) + c_{01}P(H_1|y)$$

$$(c_{10} - c_{00})P(H_0|y) \leq (c_{01} - c_{11})P(H_1|y).$$

Now since we prefer the correct solution to an error, $c_{10} > c_{00}$ and $c_{01} > c_{11}$:

$$\frac{P(H_1|y)}{P(H_0|y)} \geq \frac{c_{10} - c_{00}}{c_{01} - c_{11}}.$$

Using Baye's Rule, H_1 is picked if and only if:

$$L(y) \geq \frac{(c_{01} - c_{11})P(H_1)}{(c_{10} - c_{00})P(H_0)} = \eta$$