CS4770

Pattern Recognition

Bayesian Decision Theory

P. J. Narayanan

Monsoon 2005

Adaptive Amusement Park

- An amusement park wants to adapt its games and rides to each visitor.
- Determine the gender of the visitor on entry by measuring some features. ω is either ω_1 (for male) or ω_2 (for female).
- From demographics: Probability $P(\omega = \omega_1) = P(\text{male}) = 0.45$. $P(\omega = \omega_2) = P(\text{female}) = 0.55$
- These are the **a priori** probabilities or **prior**s of male and female.

Guess the Gender

- A random person enters the park. What is the gender?
- **Decision Rule:**
  Decide ω_1 if $P(\omega_1) > P(\omega_2)$; else, decide ω_2.
- Females are more likely. Hence, guess is female for every visitor!
- A boring decision as the guess is same even if *know* males also will come.
- Why? Because we couldn't observe the person and have only the priors to guide us.

What if we can observe the height?

- Let x be the random variable indicating the height of the visitor, measured on entry.
- $p(x|\omega)$ gives the **class-conditional probability** density of the height x conditional on the class ω.
  It is the likelihood (or probability) of class ω generating the observation x.
- The *joint probability density* indicates when both events occur together. $p(x, \omega_i) = p(x|\omega_i)P(\omega_i) = P(\omega_i|x)p(x)$.
- **Bayes Formula:**
  $P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$, where $p(x) = \sum_i p(x|\omega_i)P(\omega_i)$

What does it say?

- $p(w_i|x)$ gives the **a posteriori** or **posterior** probability that the state is ω_i given that the height is x.

- That is, the probability of male given an observed height of 135 cm.

- Bayes formula states: $$\textbf{posterior} = \frac{\textbf{likelihood} \times \textbf{prior}}{\textbf{evidence}}$$

- Bayes formula converts a priori probability $P(\omega_i)$ to a posteriori probability $P(\omega_i|x)$ with the help of the likelihood or conditional probability $p(x|\omega_i)$.

Modified Guess

- Observe the height. What is the modified decision?
- New decision rule: ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; else ω_2.

  Or, decide ω_1 if the **likelihood ratio** $\frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$.
- The LHS depends on the measurement. It should be greater than a threshold that depends on the priors.
- At 175 cm: likelihood for male is 0.7, and for female is 0.2. Decide male as $0.7/0.2 = 3.5 > 0.55/0.45 = 1.22$.
- At 135 cm: $p(x|\omega_2) = 0.8$ and $p(x|\omega_1) = 0.3$.
  Decide ω_2 as $0.3/0.8 < 1.22$.

Error in the Decision

- What's the probability of error, given an observation x?
- $P(error|x) = P(\omega_2|x)$ if decided on $\omega_1;$ else $P(\omega_1|x)$.
- For Bayes decision rule, $P(error|x) = \min\{P(\omega_1|x), P(\omega_2|x)\}$.
- Total error obtained by integrating over all x.

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error|x)p(x)dx$$

- Bayes decision rule minimizes $P(error|x)$ for every x. Therefore, it also minimizes the total error probability $P(error)$.

Risk in Decision

- Males get angry if misclassified as a female and damage property. Females also get angry but only sulk a little.

- Park owner wants of minimize his risk and would try to avoid males getting classified as females even at the cost of the reverse.

- Let α_i be an action i and $\lambda(\alpha_i|\omega_j) = \lambda_{ij}$ the risk in taking that action from state ω_j.

- Simple case: α_i means deciding ω_i.

Two-Category Classification

- Risk involved in each decision:

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x})$$
$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x})$$

- Choose the decision with lower risk.
  Decide ω_1 if $(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x})$

- Using Bayes rule, this reduces to: Decide ω_1 if

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)}$$

- Ordinarily, $\lambda_{12} > \lambda_{11}$ and $\lambda_{21} > \lambda_{22}$.

- As λ_{12} increases, the threshold increases. Makes it more difficult to decide ω_1.

- Makes sense as risk of misclassifying as ω_1 increases with λ_{12}.

- Risk moves the threshold in favour of the less risky action.

- In our example: $\lambda_{12} = 0.2$ (patient females) while $\lambda_{21} = 0.7$ (aggressive males).

- At 135 cm: $0.3/0.8 = 0.375 > 0.2{*}0.55/(0.7{*}0.45) = 0.35$. Hence, decide male!

The General Situation

- There are c states or **categories** $\{\omega_1, \omega_2, \ldots, \omega_c\}$.
- One of a actions $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ is taken.
- Observation is a d-dimensional feature vector $\mathbf{x} = [x_1, x_2, \ldots, x_d]^\mathsf{T}$.
- $\lambda(\alpha_i|\omega_j)$: loss of taking action α_i given that the state is ω_j.
- Posteriors: $P(\omega_i|\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i) / \sum_i^c p(\mathbf{x}|\omega_i)P(\omega_i)$.
- Expected loss or **risk**: $R(\alpha_i|\mathbf{x}) = \sum_{j=1}^c \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$.
- Overall risk: $R = \int_{\mathbf{x}} R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}$.

Bayes Decision Rule under Risk

- Overall risk is minimized if the conditional risk $R(\alpha_i|\mathbf{x})$ is minimum for every $\mathbf{x}$.
- Bayes decision rule: Choose α_i that minimizes $R(\alpha_i|\mathbf{x})$
- Take action α_k (from $\alpha_1 \ldots \alpha_a$) where

$$k = \arg\min_i R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

.

- The minimum risk $R^*(\alpha_i|\mathbf{x})$ is called the **Bayes risk**.

Minimum Error-Rate Classification

- If we penalize wrong classifications equally, $\lambda_{ij} = 1 - \delta_{ij}$.
- Risk $R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda_{ij} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$.
- $R(\alpha_i|\mathbf{x})$ is minimum for the decision i for which the posterior $P(\omega_i|\mathbf{x})$ is maximum.
- Same decision rule as the Bayes classifier.
- In the two-category case, if the loss for one action is greater than the other, the regions for that action will shrink.

Discriminant Functions

- Classifiers as functions $g_i(x),\ i = 1 \ldots c$ such that:

  Classify $\mathbf{x}$ as class ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x}), \forall j \neq i$

- For the general Bayes classifier with risks,

  $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$.

- Without loss functions (or for minimum error-rates)

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)/p(\mathbf{x})$$
$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$
$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

- Many ways to satisfy the extremum condition.

Two-Category Case

- Also called the **dichotomizer**.
- Instead of defining $g_1(.)$ and $g_2(.)$, we can define
$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}).$$
- Classify $\mathbf{x}$ as ω_1 if $g(\mathbf{x}) > 0$. Else classify it as ω_2.
- The following are commonly used:
$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$
$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Error Probability

- What's the probability of error for any classifier? Deciding ω_1 when ω_2 and vice versa.
- $P(err) = P(\mathbf{x} \in \mathcal{R}_1, \omega_2) + P(\mathbf{x} \in \mathcal{R}_2, \omega_1)$
  $P(err) = \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)P(\omega_1)d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)P(\omega_2)d\mathbf{x}$
  ($P(\mathbf{x}, \omega_2) = p(\mathbf{x}|\omega_2)P(\omega_2)$. If you limit $\mathbf{x} \in \mathcal{R}_1$, this becomes the integral over $\mathcal{R}_1$)
- Error under the curve. Minimized for Bayesian classifiers.
- For multicategory case, compute probability of correctness:

$$P(corr) = \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, \omega_i) = \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|\omega_i)P(\omega_i)d\mathbf{x}$$

- Bayesian decision rule is optimal for this case as the integrand is maximum for every $\mathbf{x}$.

Abstract to Concrete

- The discussion so far has been based on an abstract probability distributions for $p(\mathbf{x}|\omega), P(\omega)$.
- Let us look at a probability distribution that is of great interest.
- What is the Bayesian Decision Boundary for a specific distribution?
- Consider the **Normal Distribution** or the **Gaussian Distribution**.

Normal Distribution

- Analytically tractable, nice properties. Hence popular.
- **Central limit theorem:** The sum of a number of independent random variables of any distribution tends to a normal distribution!
- Assume a process outcome is a sum of many independent random parameters. Together they will become a Gaussian.
- True about most “real” processes!?

Normal Distribution: $N(\mu, \sigma^2)$

- $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left[\frac{(x-\mu)}{\sigma}\right]^2}$.
- μ is the mean and σ^2 is the variance.
- Nice bell curve centered around μ.
- Only 32% beyond $\mu \pm \sigma$; Only 5% beyond $\mu \pm 2\sigma$; Only 0.3% beyond $\mu \pm 3\sigma$.

$$\int_{-\infty}^{\infty} p(x)dx = 1.0; \quad \int_{-\infty}^{\infty} x\, p(x)dx = \mu;$$

$$\int_{-\infty}^{\infty} (x-\mu)^2\, p(x)dx = \sigma^2$$

Multivariate Normal: $N(\mu, \mathbf{\Sigma})$

- $p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\mu)}$.
- $\mu = E[\mathbf{x}]$. $\mathbf{\Sigma} = E[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^{\mathsf{T}}]$
- Expectation of a vector/matrix: Compute expectations of its components independently.
- **Covariance matrix** $\mathbf{\Sigma}$ has components:
  $\sigma_{ii} = \sigma_i = E[(x_i - \mu_i)^2], \quad \sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$
  $\mathbf{\Sigma}$ is a symmetric matrix.
- If i and j are statistically independent, $\sigma_{ij} = 0$.

Hyperellipsoids

- $p(x) = c$ denotes a d-dimensional hyperellipsoid defined by $(\mathbf{x} - \mu)^{\mathsf{T}}\Sigma^{-1}(\mathbf{x} - \mu) = k$.
- The principal axes of this ellipsoid are the eigenvectors of $\boldsymbol{\Sigma}$. Eigenvalues give their lengths.
- When different components of the vector have different extents (that is, variances), the distance between two vectors is skewed in favour of some.
- $r^2 = (\mathbf{x} - \mu)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu)$ gives a normalized distance to the mean. Called the **Mahalanobis distance**.

Transforming Vectors

- If $\mathbf{A}$ is a $d \times k$ matrix. $\mathbf{y} = \mathbf{A}^\intercal \mathbf{x}$ is a k-vector which is the transformation of $\mathbf{x}$.
- If $\mathbf{x}$ is normally distributed as $N(\mu, \mathbf{\Sigma})$, $\mathbf{y}$ is normally distributed as $N(\mathbf{A}^\intercal \mu, \mathbf{A}^\intercal \mathbf{\Sigma} \mathbf{A})$.
- For $\mathbf{A}_\mathbf{w} = \mathbf{\Phi} \mathbf{\Lambda}^{-\frac{1}{2}}$ where $\mathbf{\Phi}$ is a matrix formed with the eigenvectors of $\mathbf{\Sigma}$ as its columns and $\mathbf{\Lambda}$ is a diagonal matrix of the eigenvalues, $\mathbf{A}_\mathbf{w}\mathbf{x}$ is a transformation that converts the hyperellipsoid to a hypersphere. (That is, $\mathbf{\Sigma} = \mathbf{I}$ after transform).
- $\mathbf{A}_\mathbf{w}$ is called a **whitening transform**.

Discriminant Functions for Gaussians

- The likelihoods $p(\mathbf{x}|\omega)$ are Normal distributed.

- Let us use the definition

$$
\begin{aligned}
\mathbf{g}_i(\mathbf{x}) &= \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \\
&= -\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{i}})^{\mathsf{T}} \mathbf{\Sigma}_i^{-1} (\mathbf{x} - \mu_{\mathbf{i}}) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln P(\omega_i)
\end{aligned}
$$

- The second term is a constant and plays no role. Only the first term depneds on $\mathbf{x}$ or the sample!.

Simple Case: Equal, Independent Variances

- Let $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$.
- Samples from each cluster fall in hyperspherical clusters centered around their respective means $\mu_{\mathbf{i}}$.
- $\mathbf{g}_i(\mathbf{x}) = -\frac{||\mathbf{x} - \mu_i||^2}{2\sigma^2} + \ln P(\omega_i)$.
- With equal priors, distance to μ_i alone matters.
- Otherwise, priors plus a quantity proportional to the squared Mahalanobis distance is the function for each class.

- With independent and unequal variances, $\mathbf{g}_i(\mathbf{x}) = \ln P(\omega_i) - \sum_j \frac{||x_j - \mu_{ij}||^2}{2\sigma_i^2}$.
- The components with larger variance matter less.

Linear Form

- Expanding the equation, we get:
  $\mathbf{g}_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}^\intercal\mathbf{x} - 2\mu_i^\intercal\mathbf{x} + \mu_i^\intercal\mu_i) + \ln P(\omega_i)$.

- Since $\mathbf{x}^\intercal\mathbf{x}$ is same for all classes, the discriminant function is linear in $\mathbf{x}$.

- $\mathbf{g}_i(\mathbf{x}) = \mathbf{w}_i^\intercal\mathbf{x} + w_{i0}$, where the w_{i0} is a bias or threshold.
  $\mathbf{w}_i = \frac{\mu_i}{\sigma^2}$ and $w_{i0} = \ln P(\omega_i) - \frac{\mu_i^\intercal\mu_i}{2\sigma^2}$

- A **linear machine** when the discriminant functions are linear.

Between classes i and j

- $\mathbf{g_i}(\mathbf{x}) = \mathbf{g_j}(\mathbf{x})$ when $\mathbf{w}^\intercal(\mathbf{x} - \mathbf{x_0}) = \mathbf{0}$ where $\mathbf{w} = \mu_\mathbf{i} - \mu_\mathbf{j}$ and $\mathbf{x}_0 = \frac{\mu_i+\mu_j}{2} - \frac{\sigma^2}{||\mu_i-\mu_j||^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$
- $\mathbf{g}_i$ gives a hyperplane through $\mathbf{x_0}$, perpendicular to $\mu_i - \mu_j$.
- With equal priors, $\mathbf{x_0}$ is at the centre of the line joining μ_i and μ_j. Otherwise, it gets shifted away from the class with higher prior.
- With low σ^2, the effect of priors is small.
- If all priors are equal, this becomes a **minimum distance classifier**.

Identical Σ

- Discriminant function reduces to
$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_\mathbf{i})^\mathsf{T}\mathbf{\Sigma}^{\text{-1}}(\mathbf{x} - \mu_\mathbf{i}) + \ln P(\omega_i)$

- With equal priors, closest class using Mahalanobis distance.

- Can eliminate the $\mathbf{x^2}$ term from the expansion.

- Linear in $\mathbf{x}$ with $\mathbf{g}_i(\mathbf{x}) = \mathbf{w}_i^\mathsf{T}\mathbf{x} + w_{i0}$ where $\mathbf{w}_i = \mathbf{\Sigma}^{\text{-1}}\mu_i$ and $w_{i0} = -\frac{1}{2}\mu_i^\mathsf{T}\mathbf{\Sigma}^{\text{-1}}\mu_i + \ln P(\omega_i)$.

- Decision boundaries between classes are hyperplanes

- Between two classes, discriminant function becomes $\mathbf{w}^{\mathsf{T}}(\mathbf{x}-\mathbf{x_0})=\mathbf{0}$ where $\mathbf{w}=\boldsymbol{\Sigma}^{\text{-1}}(\mu_{\mathbf{i}}-\mu_{\mathbf{j}})$ and $\mathbf{x_0}=\frac{\mathbf{1}}{\mathbf{2}}(\mu_{\mathbf{i}}+\mu_{\mathbf{j}})-\frac{\ln\mathbf{P}(\omega_{\mathbf{i}})-\ln\mathbf{P}(\omega_{\mathbf{j}})}{(\mu_{\mathbf{i}}-\mu_{\mathbf{j}})^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathbf{i}}^{\text{-1}}(\mu_{\mathbf{i}}-\mu_{\mathbf{j}})}(\mu_{\mathbf{i}}-\mu_{\mathbf{j}})$

- Hyperplane separating i and j are not perpendicular to the line joining the means as $\Sigma^{\text{-1}}(\mu_i-\mu_j)$ is not in the direction of $\mu_i-\mu_j$.

- The hyperplane does go through $\mathbf{x_0}$. With equal priors, $\mathbf{x_0}$ is midway between the means. Otherwise, it is shifted away from the more likely class.

Arbitrary Σ_i

- In the general case, $g_i(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^\mathsf{T}\mathbf{x} + w_{i0}$
  where $\mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}$, $\mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\mu_i$ and
  $w_{i0} = -\frac{1}{2}\mu_\mathbf{i}^\mathsf{T}\mathbf{\Sigma}_\mathbf{i}^{-1}\mu_\mathbf{i} - \frac{1}{2}\ln|\mathbf{\Sigma}_\mathbf{i}| + \ln P(\omega_i)$

- Between two classes $g(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{W}_{ij}\mathbf{x} + \mathbf{w}_{ij}^\mathsf{T}\mathbf{x} + w_{ij0}$

- Decision boundaries are **hyperquadrics**.
  In 2-D, $g_i(\mathbf{x}) = ax^2 + (b+c)xy + dy^2 + ex + fy + k$, a general quadratic that could be a parabola, ellipse, hyperbola (even a line!)

- With multiple classes, decision boundaries are hyperquadrics; the feature space is divided by intersecting hyperquadrics, generating complex boundaries.

- Even in one dimension, the classes may not be simply connected with two classes!

- $ax^2 + bx + c = 0$ has 2 solutions in general. Between them lies one region and beyond in either direction the other region.

Simple 2-Class Case: Signal Detection

- Looking for a signal (ω_2) when the default is ω_1.
- $p(\mathbf{x}|\omega_1) \sim N(\mu_1, \sigma), p(\mathbf{x}|\omega_2) \sim N(\mu_2, \sigma)$,
- **Discriminability:** $d = \frac{|\mu_1 - \mu_2|}{\sigma}$.
- Higher the discriminability, better the performance.
- Classifier uses a threshold x^*. $\omega_2 : x > x^*$.
- Measures the 2-class classifier by how samples from different classes are decided.

Hit, Miss, False Alarm, Correct Rejection

- **Hit**: Correct acceptance.
  Probability = $P(R_2|\omega_2) = p(x > x^*|x \in \omega_2)$.
- **False alarm**: Wrong acceptance.
  Probabilty = $P(R_2|\omega_1) = p(x > x^*|x \in \omega_1)$
- **Miss**: Wrong rejection.
  Probability = $P(R_1|\omega_2) = p(x < x^*|x \in \omega_2)$
- **Correct Rejection**: Correct rejection.
  Probability = $P(R_1|\omega_1) = p(x < x^*|x \in \omega_1)$

Receiver Operating Characteristic (ROC)

- Hit vs false alarm curve for varying x^*.
- Exactly one curve for each discriminability value d.
- Curve becomes steep (closer to hit axis) as d increases.
- For multidimensional case, such a curve when a single decision parameter is changed is called the operating characteristic.

For Discrete Features

- Integrals are replaced by summations when features are discrete.
- Bayes theorem is similar, but using probabilities instead of densities.
- $P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{\Sigma_j P(\mathbf{x}|\omega_j)P(\omega_j)}$
- Conditional risk $R(\alpha_i|\mathbf{x})$ is as before. Select the action for minimum risk as $\arg\min_i R(\alpha_i|\mathbf{x})$

Independent Binary Features

- Independent features with binary values. $x_i = \pm 1$ or $0/1$.

- For a 2-class problem, let
  $p_i = P(x_i = 1 | \omega_1)$ and $q_i = P(x_i = 1 | \omega_2)$.

- Likelikhoods are: $P(\mathbf{x}|\omega_1) = \Pi_i p_i^{x_i}(1 - p_i)^{1-x_i}$ and $P(\mathbf{x}|\omega_2) = \Pi_i q_i^{x_i}(1 - q_i)^{1-x_i}$.

- The discriminant function using likelihood ratios is

$$g(\mathbf{x}) = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

- Can be written as $g(\mathbf{x}) = \sum_i w_i x_i + w_0$ where,
  $w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)}$ and $w_0 = \sum_i \ln \frac{(1-p_i)}{(1-q_i)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$

- Possible values of $\mathbf{x}$ are on the vertices of a d-dimensional hypercube. The decision surface is a hyperplane that separates the vertices belonging to each class.

- If $p_i > q_i$, w_i is positive. x_i contributes positively to ω_1. Otherwise, negatively!

- Priors introduce a bias in favour of the respective classes.

Packing Fruits

- Mangoes are packed and shipped to China. Oranges are packed off to Calcutta.
- Chinese are picky. If an orange is found in the mango box, you are fined 100 rupees. If a mango is found in the box of oranges, Calcuttans fine you 10 rupees.
- The likelihood ratio needs to be greater than $\frac{100}{10}\frac{1/3}{2/3} = 5$ to be decided as a mango in our example.
- For green, likelihood ratio is 0.7 / 0.2 = 3.5. Hence, decide orange, since you don't want to risk much! Decide orange even if the likelihoods are 0.95 and 0.2.

Sour Fruits and Picky Customers

- Picked fruits are sent to Delhi, Hyderabad or Chennai.
- Loss values for sending mangoes to these cities are: $3.0, 2.0$, and 1.0. The losses for sending oranges are: $1.0, 3.0$, and 5.0.
- For an observed green: $[P(g|M) = 0.7, P(g|O) = 0.2]$

  $P(M|g) = 1.4k, P(O|g) = 0.2k$

  $R(D|g) = \lambda_{DM} P(M|g) + \lambda_{DO} P(O|g) = k(3 * 1.4 + 1 * 0.2) = 4.4k$

  $R(H|g) = \lambda_{HM} P(M|g) + \lambda_{HO} P(O|g) = k(2 * 1.4 + 3 * 0.2) = 3.4k$

  $R(C|g) = \lambda_{CM} P(M|g) + \lambda_{CO} P(O|g) = k(1 * 1.4 + 5 * 0.2) = 2.4k$
- Choose the action with minimum risk. Send to Chennai!

- For yellow-red: $[P(y|M) = 0.2, P(y|O) = 0.8]$

  $P(M|y) = 0.4k, P(O|y) = 0.8k$

  $R(D|y) = \lambda_{DM} P(M|y) + \lambda_{DO} P(O|y) = k(3 * 0.4 + 1 * 0.8) = 2.0k$

  $R(H|y) = \lambda_{HM} P(M|y) + \lambda_{HO} P(O|y) = k(2 * 0.4 + 3 * 0.8) = 3.2k$

  $R(C|y) = \lambda_{CM} P(M|y) + \lambda_{CO} P(O|y) = k(1 * 0.4 + 5 * 0.8) = 4.4k$

- Destination: Delhi!

- If conditional probabilities are equal (0.5, 0.5), posteriors are 1.0 and 0.5. And risks are equal at $3.5k$.

- If posteriors are equal, risks are $4k, 5k, 6k$. Choose Delhi!