3. Bayes and Normal Models

Aleix M. Martinez
aleix@ece.osu.edu

Why Bayesian?

• If all our research (in PR) was to disappear and you could only save one theory, which one would you save?
• Bayesian theory is probably the most important one you should keep.
• It’s simple, intuitive and optimal.
• Reverend Bayes (1763) and Laplace (1812) set the foundations of what we now know a Bayes theory.

State of nature: class

• A sample (usually) corresponds to a state of nature; e.g. salmon and sea bass.
• The state of nature usually corresponds to a set of discreet categories (classes). Note that the continuous also exists.
• Priors: some classes might occur more often or might be more “important,” \( P(w_i) \).

Decision rule

• We need a decision rule to help us determine to which class a testing vector belongs.
• Simplest (useless): \( C = \arg \max_i p(w_i) \)
• Posterior probability: \( p(w_i \mid x) \), \( x \) is the (observed) data.
• Obviously, we do not have \( p(w_i \mid x) \).
• But we can estimate: \( p(x \mid w_i) \) & \( P(w_i) \).

Bayes Theorem (yes, the famous one)

\[
p(w_i \mid x) = \frac{p(x \mid w_i)P(w_i)}{p(x)}
\]

\[
p(x) = \sum_{i=1}^{c} p(x \mid w_i)P(w_i)
\]

• Bayes decision rule: \( \arg \max_i p(w_i \mid x) \)

Rev. Thomas Bayes (1702-1761)

• During his lifetime, Bayes was a defender of Isaac Newton’s calculus, and developed several important results of which the Bayes Theorem is his most known and, arguably, most elegant. This theorem and the subsequent development of Bayesian theory are among the most relevant topics in pattern recognition and have found applications in almost every corner of the scientific world. Bayes himself did not, however, provide the derivations of the Bayes Theorem as this is now known to us. Bayes develop the method for uniform priors. This result was later extended by Laplace and contemporaries. Nonetheless, Bayes is generally acknowledge as the first to have established a mathematical basis for probability inference.
Multiple random variables

• To be mathematically precise, one should write \( p_x(x | w_j) \) instead of \( p(x | w_j) \), because this probability density function depends on a single random variable \( X \).

• In general there is no need for distinction (e.g., \( p_X \) & \( p_Y \)). Shall this arise, we will use the above notation.

(see Appendix A.4)

Loss function & decision risk

• States exactly how costly each action is, and is used to convert a probability determination into a decision.

\[
\text{classes: } \{ w_1, \ldots, w_p \} \quad \text{actions: } \{ \alpha_1, \ldots, \alpha_n \}
\]

\[
\text{loss function: } \lambda(\alpha_j | w_j) \quad \text{the cost (risk) of going from } w_j \text{ to } \alpha_j
\]

Conditional Risk:
\[
R(\alpha_j | x) = \sum_{j=1}^{p} \lambda(\alpha_j | w_j) p(w_j | x)
\]

Bayes decision rule

Conditional risk:
\[
R(\alpha_i | x) = \sum_{j=1}^{p} \lambda(\alpha_j | w_j) p(w_j | x)
\]

Bayes decision rule (Bayes risk):
\[
\arg \min_{i} R(\alpha_i | x)
\]

• The resulting minimum overall risk is called the Bayes risk.

A simple example

• Two-class classifier:

\[
R(\alpha_1 | x) = \lambda_{11} p(w_1 | x) + \lambda_{12} p(w_2 | x)
\]

\[
R(\alpha_2 | x) = \lambda_{21} p(w_1 | x) + \lambda_{22} p(w_2 | x)
\]

• Decision rule: \( \alpha_i \) (or \( w_i \)) if \( R(\alpha_1 | x) < R(\alpha_2 | x) \), \( (\lambda_{11} - \lambda_{12}) p(w_1 | x) > (\lambda_{21} - \lambda_{22}) p(w_2 | x) \)

• Applying Bayes:

\[
\frac{p(x | w_1)}{p(x | w_2)} > \frac{\lambda_{11} - \lambda_{21}}{\lambda_{21} - \lambda_{11}} \frac{P(w_1)}{P(w_2)}
\]

Feature Space: Geometry

• When \( x \in \mathbb{R}^d \), we have our d-dimensional feature space.

• Sometimes, this feature space is considered to be an Euclidean space; but as we’ll see many other alternatives exists.

• This allows for the study of PR problems form a geometric point of view. This is key to many algorithms.
**Discriminant functions**

- We can construct a set of discriminant functions: \( g_i(x) \), \( i = 1, \ldots, c \).
- We classify a feature vector as \( w_j \) if:
  \[ g_i(x) > g_j(x), \forall j \neq i \]
- The Bayes classifier is: \( g_i(x) = -R(\alpha_i | x) \).
- If errors are to be minimized, one needs to minimize the probability of error:
  \[ \lambda(\alpha_i | w_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \]

- If we use Bayes & minimum-error-rate classification, we get:
  \[ g_i(x) = p(w_i | x) = \frac{p(x | w_i)P(w_i)}{\sum_{j} p(x | w_j)P(w_j)} \]

- Sometime we find more convenient to write this equation as: \( g_i(x) = \ln p(x | w_i) + \ln P(w_i) \).
- Geometry (key point): the goal is to divide the feature space into \( c \) decision regions, \( R_1, \ldots, R_c \).
- Classification is also known as hypothesis testing.

**Other criterion**

- In some applications the priors are not known.
- In this case, we usually attempt to minimize the worst overall risk.
- Two approaches for that are: the minimax and the Neyman-Pearson criteria.
Normal Distributions & Bayes

- So far we have used $p(x|w)$ and $p(w)$ to specify the decision boundaries of a Bayes classifier.
- The Normal distribution is the most typical PDF for $p()$.
- Recall the central limit theorem.

Central Limit theorem (simplified)

Assume that the random variables $X_1,...,X_n$ are iid, each with finite mean and variance. When $n \to \infty$, the standardized random variable converges to a normal distribution.

(see Stark & Woods pp. 225-230)

Univariate case

- The Gaussian distribution is:
  
  $$p(x) = \frac{1}{\sqrt{2\pi \sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  $\mu = \int xp(x)dx$

  $\sigma^2 = \int (x-\mu)^2 p(x)dx$

Multivariate case ($d>1$)

$$p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}$$
Distances

• The general distance in a space is given by:
  \[ d^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \]
where \( \Sigma \) is the covariance matrix of the distribution (or data).
• If \( \Sigma = I \) then the above equation becomes the Euclidean (norm 2) distance.
• If \( \Sigma \) is Normal, this distance is called \textit{Mahalanobis} distance.

Example (2D Normals)

Moments of the estimates

• In statistics the estimates are generally known as the \textit{moments} of the data.
• The first moment is the \textit{sample} mean.
• The second, the \textit{sample} autocorrelation matrix:
  \[ \hat{S} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T. \]

Central moments

• The variance and the covariance matrix are special cases, because they depend on the mean of the data which is unknown.
• Usually we solve that by using the sample mean:
  \[ \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T. \]
• This is the sample covariance matrix.

Whitening Transformation

• Recall, it is sometime convenient to represent the data in a space where its sample covariance matrix equals the identity matrix, \( I \).

\[ A^T \Sigma A = I \]

Linear transformations

• An n-dimensional vector \( X \) can be transformed linearly to another, \( Y \), as:
  \[ Y = A^T X \]
• The mean is then: \( M_Y = E(Y) = A^T M_X \)
• The cov.: \( \Sigma_Y = A^T \Sigma_X A \)
• The order of the distances in the transformed space is identical to the one in the original space.
**Orthonormal transformation**

- Eigenanalysis: $\Sigma \Phi = \Phi \Lambda$
- Eigenvectors: $\Phi = [\Phi_1, \ldots, \Phi_p]$
- Eigenvalues: $\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \ddots \\ 0 & \lambda_p \end{pmatrix}$
- The transformation is then: $Y = \Phi^T X$
  $$\Sigma_Y = \Phi^T \Sigma_X \Phi = \Lambda$$ (recall $\Phi^T \Phi = I$ & $\Phi^T = \Phi^*$)

**Whitening**

- To obtain a covariance matrix equal to the identity matrix we can apply the orthogonal transformation first and then normalize the result with $\Lambda^{-1/2}$:
  $$Y = \Lambda^{-1/2} \Phi^T X = \Phi^T X$$
  $$\Sigma_Y = \Phi_X^T \Sigma_X \Phi_X = \Lambda^{-1/2} \Phi_X^T \Sigma_X \Phi_X \Lambda^{1/2} = I$$

**Properties**

- Whitening transformations are not orthogonal transformations.
- Therefore, Euclidean distances are not preserved.
- After whitening, the covariance matrix is invariant to any orthogonal transformation:
  $$\Psi^T \Psi = \Psi^T = I$$

**Simultaneous diagonalization**

- It is usually the case where two or more covariance matrices need to be diagonal.
- Assume $\Sigma_1$ and $\Sigma_2$ are two covariance matrices.
- Our goal is to have $A^T \Sigma_1 A = I$ and $A^T \Sigma_2 A = \Lambda_2$.
- *Homework:* find the algorithm.
**Some advantages**

- Algorithms usually become much simpler after diagonalization or whitening.
- The general distance $\Rightarrow$ a simple Euclidean distance.
- Whitened data is invariant to other orthogonal transformations.
- Some algorithms require whitening to have certain properties (we’ll see this latter in the course).

**Discriminant Functions for Normal PDFs**

- The discriminant function, $g_i(x) = \ln p(x | w_i) + \ln p(w_i)$, for the Normal density, $N(\mu, \Sigma)$, is:

\[
g_i(x) = -\frac{1}{2}(x - \mu)^T \Sigma_i^{-1} (x - \mu) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln p(w_i)
\]

- Possible scenarios (or assumptions):
  - Sometimes, we might be able to assume $\Sigma_i = \sigma^2 I$.
  - A more general case is when all covariance matrices are identical; i.e. homoscedastic.
  - The most complex case is when $\Sigma_i = \text{arbitrary}$; that is, heteroscedastic.

**Homoscedastic:** $\Sigma_i = \Sigma$

**Mahalanobis**

\[
g_i(x) = -\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu) + \ln p(w_i)
\]
**Heteroscedastic**

\[ \Sigma_i = \text{arbitrary} \]

- In the 2-class case, the decision surface is an hyperquadric (e.g. hyperplanes, hyperspheres, hyperhyperboloids, etc.).
- These decision boundaries may not be connected.
- Any hyperquadric can be given (represented) by two Gaussian distributions.

\[ \Sigma_i = \text{arbitrary} \]

**Project #1**

1. Implement these three cases using Matlab (see pp. 36–41 for details). 2D and/or 3D plots.
2. Generalize the algorithm to more than two classes – Gaussians.
3. Simulate different Gaussians and distinct priors.

**Bayes Is Optimal**

- If our goal is to minimize the classification error, then Bayes is optimal (you cannot do better than Bayes – ever).
- In general, if \( p(x \mid w_1)P(w_1) > p(x \mid w_2)P(w_2) \), it is preferable to classify \( x \) in \( w_1 \) so that the smallest integral contributes to the error (see next slide) \( \Rightarrow \) This is what Bayes does.
- There is no possible smaller error.

\[
P(\text{error}) = P(x \in R_{w_1}) + P(x \in R_{w_2}) = \\
= \int_{R_{w_1}} p(x \mid w_1)P(w_1)dx + \int_{R_{w_2}} p(x \mid w_2)P(w_2)dx
\]
The multiclass case:

\[
P(\text{correct}) = \sum_{i=1}^{C} P(x \in R_i | w_i) = \sum_{i=1}^{C} \int p(x | w_i) P(w_i) \, dx.
\]

- Bayes yields the smallest error. But which is the actual error? The above equation cannot be readily computed, because the regions \( R_i \) may be very complex.

### Error Bounds: How to calculate the error?

- Several approximations are easier to compute (usually upper bounds):
  - Chernoff bound.
  - Bhattacharyya bound (assumes pdf are homoscedastic).
- These bounds can only be applied to the 2-class case only.

#### Chernoff Bound

- For this, we need an integral eq. that we can solve. For example,
  \[
P(\text{error}) = \int P(\text{error} \mid x) p(x) \, dx,
\]
  where
  \[
P(\text{error} \mid x) = \begin{cases} P(w_1 \mid x), & \text{if we decide } w_1 \\ P(w_2 \mid x), & \text{if we decide } w_2. \end{cases}
\]

\[
\int p'(x \mid w_i) p^{-1}(x \mid w_j) \, dx = e^{-k(s)},
\]

where

\[
k(s) = \frac{s(1-s)}{2} (\mu_2 - \mu_1)^T (\Sigma_1 + (1-s)\Sigma_2)^{-1} (\mu_2 - \mu_1) + \frac{1}{2} \ln \frac{\det(\Sigma_1 + (1-s)\Sigma_2)}{\det\Sigma_1}.
\]

#### Bhattacharyya Bound

- When the data is homoscedastic, \( \Sigma_i = \Sigma_z \), the optimal solution is \( s=1/2 \).
- This is the Bhattacharyya bound.
- A tighter bound is the asymptotic nearest neighbor error, which is derived from:
  \[
p(\text{error}) \leq 2 \int \frac{p(x \mid w_i) p(x \mid w_j)}{p(x)} p(w_i) \, dx \leq \int \sqrt{p(x \mid w_i) p(x \mid w_j)} p(w_i) \, dx.
\]
Closing Notes

- Bayes is important because it minimizes the probability of error. In that sense we say it’s \textit{optimal}.
- Unfortunately, Bayes \textit{assumes} that the conditional densities and priors are \textit{known} (or can be estimated); which is not necessarily true.
- In general, not even the form of these probabilities is known.
- Most PR approaches attempt to solve these shortcomings. This is, in fact, what most of PR is all about.

On the + side: a simple example

- We want to predict whether a student will pass or not a test.
- \(Y=1\) denotes pass, \(Y=0\) failure.
- The observation is a single random variable \(X\) which specifies the hours of study.
- Let \(p(Y=1|X=x) = \frac{x}{c+x}\). Then:
  \[
  g(x) = \begin{cases} 
  1 & \text{if } p(Y=1|X=x) > 1/2; \text{ i.e. } x > c \\
  0 & \text{otherwise}
  \end{cases}
  \]

Optional homework

- Using Matlab generate \(n\) observations of \(P(Y=1|X=x)\) and \(P(Y=0|X=x)\).
- Approximate each using a Gaussian distribution.
- Calculate the Bayes decision boundary and classification error.
- Select several arbitrary values for \(c\) and see how well you can approximate them.

Hints

- Error = \(\min \{P(Y=1|X=x), P(Y=0|X=x)\}\).
- Plot the original distribution to help you.