# Naïve Bayesian Classifier (method review)

Grishin Andrey (group: M05-318a)

Department of Machine Learning and Digital Humanities

Moscow Institute of Physics and Technology

November 11, 2023

# Plan

1. Introduction
   - What is Bayesian Classifier?
   - Why (what for) do we need it?
2. Maths of Bayesian Classifier
   - Model assumptions
   - Main maths
3. Iris classification problem
   - Problem state
   - Fitting results
4. Summary
   - References
   - Conclusion

# What is Bayesian Classifier?

Bayesian Classifier — **probabilistic** classification model:

- **Foundation:** Bayes's theorem.
- **Authors:** Thomas Bayes, Richard Price.
- **Title:** "An Essay towards Solving a Problem in the Doctrine of Chances"[1].
- **Initial use:** Estimate distribution parameters.
- **Year:** 1763

[1] Reference to the article [Bayes and Price, 1763].

# Why (what for) do we need it?

- **Classifier:**
  - **Segregate** data observations.
  - **Baseline** classification models.
- **Theorem:**
  - **Estimate** distribution parameters (normal, binomial, etc.).
  - **Estimate** posterior probabilities of events.

# Model assumptions

**Let** $X \in \mathbb{R}^{n \times m}$, $y \in \{C_1, C_2, \dots, C_k\}$. **Then** need to find $f(\cdot)$ such as:

$$\mathbb{E}(y \mid X) = f(X) \iff y = f(X) + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right) \tag{1}$$

$f(\cdot) =$ Bayesian Classifier. $X_j \in \mathbb{R}^{1 \times m} : j = \overline{1, n}$ is r.v. $\mu$ and $\Sigma$ estimation – tough.

**Assume**:

- $X_j \sim \mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^{m \times 1}, \ \Sigma \in \mathbb{R}^{m \times m}$.
- $X_j = [\xi_1, \dots, \xi_m] : \xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mu_i, \sigma_i^2\right) \Rightarrow \Pr[X_j] = \prod_{i=1}^{m} \Pr[\xi_i]$.

**Note**: classification $\approx$ regression: probability $\in [0, 1]$.

# Main maths I

**Bayesian theorem** is formalized like this:

$$\Pr\left[A \mid B\right] = \frac{\Pr\left[B \mid A\right] \times \Pr\left[A\right]}{\Pr\left[B\right]} \tag{2}$$

**Bayesian classifier**, where $c_t : t = \overline{1, k}$ is the class number:

$$\Pr\left[y_j = c_t \mid X_j\right] = \frac{\Pr\left[X_j \mid y_j = c_t\right] \times \Pr\left[y_j = c_t\right]}{\Pr\left[X_j\right]} \tag{3}$$

The formula (3) can be respectively expressed as:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} \tag{4}$$

# Main maths II

**But** evidence $\Pr[X_j]$ is the same $\forall c_t : t = \overline{1, k}$. Hence, we do not strictly need it.

$$\Pr[y_j = c_t \mid X_j] \propto \Pr[X_j \mid y_j = c_t] \times \Pr[y_j = c_t] \tag{5}$$

**After expansion** of the (5) we almost have the final formula:

$$\Pr[y_j = c_t \mid X_j] \propto \left( \prod_{i=1}^{m} \Pr[X_{ji}] \right) \times \Pr[y_j = c_t] \tag{6}$$

**So**, (6) gives us the expression for $X_j$'s class:

$$\hat{C}_j = \arg \max_{t=\overline{1,k}} \left\{ \left( \prod_{i=1}^{m} \Pr[X_{ji}] \right) \times \Pr[y_j = c_t] \right\} \tag{7}$$

# Main maths III

**Q:** What exactly is $\prod_{i=1}^{m} \Pr\left[X_{ji}\right]$ and how to estimate $\left(\mu_j, \sigma_j^2\right)$?

**A:** To make things clear, let's reduce the number of **features** and **classes** to 1.

**Assumed** $\mathcal{N}\left(\mu, \sigma^2\right) \Rightarrow$ 2 parameters to be estimated. MLE method [Wilks, 1938].

$$L = \prod_{j=1}^{n} \Pr\left[X_j \mid \mu, \sigma\right] \equiv \prod_{j=1}^{n} L\left(\mu, \sigma \mid X_j\right) \tag{8}$$

No doubts to get the following optimization problem:

$$\hat{\mu}, \hat{\sigma} = \arg\max_{\mu, \sigma} \left\{ \prod_{j=1}^{n} L\left(\mu, \sigma \mid X_j\right) \right\} : \mu \in \mathbb{R}, \quad \sigma \in \mathbb{R}_{+} \tag{9}$$

# Main maths IV

**Apply** $\ln(x)$ as extrema is invariant to monotonic increasing transforms.

$$\arg\max_{\mu,\sigma}\left\{\prod_{j=1}^{n} L\left(\mu,\sigma \mid X_j\right)\right\} \sim \arg\max_{\mu,\sigma}\left\{\ln\left(\prod_{j=1}^{n} L\left(\mu,\sigma \mid X_j\right)\right)\right\} \tag{10}$$

It leads to simplification:

$$\arg\max_{\mu,\sigma}\left\{\sum_{j=1}^{n} \ln\left(L\left(\mu,\sigma \mid X_j\right)\right)\right\} \tag{11}$$

**Finally** to estimate $\mu$ and $\sigma$ we have to optimize $L(\mu,\sigma) = \sum_{j=1}^{n} \ln\left(\Pr\left[X_j \mid \mu,\sigma\right]\right)$.

# Main maths V

And the result is the following (**unbiased** variance):

$$\max_{\mu,\sigma}\left\{\sum_{j=1}^{n}\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{(X_j-\mu)^2}{2\sigma^2}\right)\right)\right\}\Rightarrow\begin{cases}\hat{\mu} & = & \dfrac{\sum_{j=1}^{n}X_j}{n}\\[2ex] \hat{\sigma}^2 & = & \dfrac{\sum_{j=1}^{n}\left(X_j-\bar{X}\right)^2}{n-1}\end{cases}\tag{12}$$

# Main maths VI

With $\boldsymbol{m}$ i.i.d. features we obtain the following estimation ($i = \overline{1, m}$):

**Mean estimation**

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} X_{ji}}{n} \tag{13}$$

**Variance estimation**

$$\hat{\sigma}_i^2 = \frac{\sum_{j=1}^{n} \left(X_{ji} - \bar{X}_i\right)^2}{n-1} \tag{14}$$

# Main maths VII

With $\boldsymbol{m}$ features and $\boldsymbol{k}$ classes estimation (13) (14) turns into ($i = \overline{1, m}$, $t = \overline{1, k}$):

**Mean estimation**

$$\hat{\mu}_i \left(c_t\right) = \frac{\sum_{j=1}^{n} X_{ji} \left[y_j = c_t\right]}{\sum_{j=1}^{n} \left[y_j = c_t\right]} \tag{15}$$

**Variance estimation**

$$\hat{\sigma}_i^2 \left(c_t\right) = \frac{\sum_{j=1}^{n} \left(X_{ji} - \hat{\mu}_i \left(c_t\right)\right)^2 \left[y_j = c_t\right]}{\left(\sum_{j=1}^{n} \left[y_j = c_t\right]\right) - 1} \tag{16}$$

# Main maths VIII

For **simplicity**, Gaussian NB works if $\left(X_{ji}\left(c_t\right)\right.$: $\boldsymbol{j}$ obs., $\boldsymbol{i}$ feat., $\boldsymbol{c_t}$ class):

**Naïve Bayesian classifier requirement**

$$\forall c_t, i : \begin{cases} t = \overline{1, k} \\ i = \overline{1, m} \end{cases} \exists \; \hat{\mu}_i\left(c_t\right) \wedge \hat{\sigma}_i^2\left(c_t\right) \hookrightarrow X_{ji}\left(c_t\right) \sim \mathcal{N}\left(\hat{\mu}_i\left(c_t\right), \hat{\sigma}_i^2\left(c_t\right)\right) \tag{17}$$

# Main maths IX

Multiple distributions can be used:

- Normal, Binomial.
- Laplacian, Exponential.
- Rayleigh's distribution.
- etc.

# Problem state I

Obtained $X \in \mathbb{R}^{150\times4}$, $y \in \{0,1,2\}$. Classification problem: iris dataset.

| **Code** | **Name** |
|---|---|
| 0 | setosa |
| 1 | versicolor |
| 2 | virginica |

Table 1: Iris encoding

| **Code** | **Name** | **Measure** |
|---|---|---|
| 0 | sepal length | cm |
| 1 | sepal width | cm |
| 2 | petal length | cm |
| 3 | petal width | cm |

Table 2: Feature names

| **Class** | **# obs.** |
|---|---|
| setosa | 50 (0.3) |
| versicolor | 50 (0.3) |
| virginica | 50 (0.3) |

Table 3: Class balancing

# Problem state II

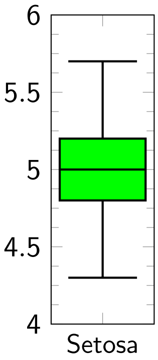



Figure 1: **Sepal length**

| **Name** | **Value** |
|---|---|
| **mean** | 5.01 |
| **std** | 0.35 |
| **max** | 5.80 |
| 75**%** | 5.20 |
| 50**%** | 5.00 |
| 25**%** | 4.80 |
| **min** | 4.30 |

Table 4: **Stats.**

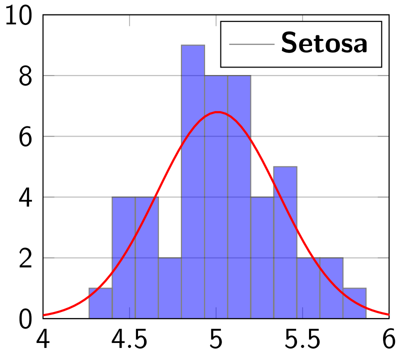



Figure 2: **Distribution**

# Problem state III

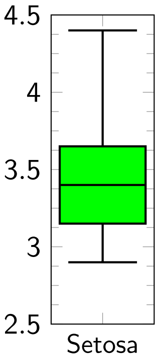



Figure 3: **Sepal width**

| Name | Value |
|---|---|
| **mean** | 3.43 |
| **std** | 0.38 |
| **max** | 4.40 |
| 75**%** | 3.68 |
| 50**%** | 3.40 |
| 25**%** | 3.20 |
| **min** | 2.30 |

Table 5: **Stats.**

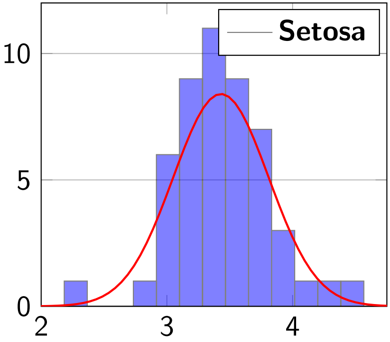



Figure 4: **Distribution**

# Problem state IV

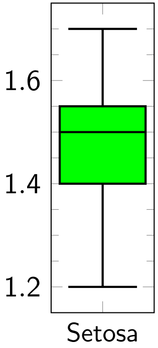



Figure 5: **Petal length**

| **Name** | **Value** |
|---|---|
| **mean** | 1.46 |
| **std** | 0.17 |
| **max** | 1.90 |
| 75**%** | 1.58 |
| 50**%** | 1.50 |
| 25**%** | 1.40 |
| **min** | 1.00 |

Table 6: **Stats.**

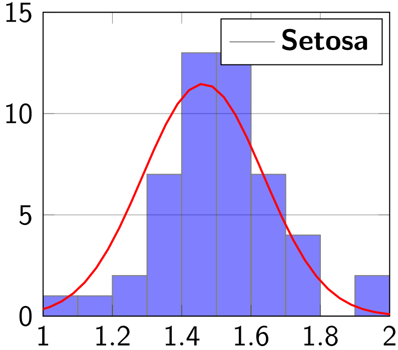



Figure 6: **Distribution**

# Problem state V

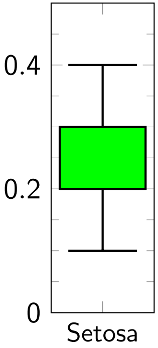



Figure 7: **Petal width**

| **Name** | **Value** |
|---|---|
| **mean** | 0.25 |
| **std** | 0.11 |
| **max** | 0.60 |
| 75**%** | 0.30 |
| 50**%** | 0.20 |
| 25**%** | 0.20 |
| **min** | 0.10 |

Table 7: **Stats.**

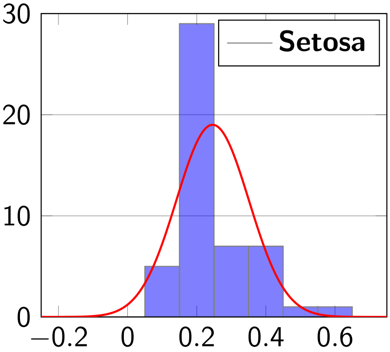



Figure 8: **Distribution**

# Problem state VI

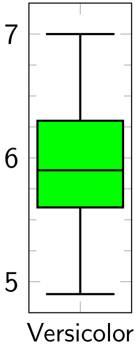



Figure 9: **Sepal length**

| Name | Value |
|---|---|
| **mean** | 5.94 |
| **std** | 0.52 |
| **max** | 7.00 |
| 75**%** | 6.30 |
| 50**%** | 5.90 |
| 25**%** | 5.60 |
| **min** | 4.90 |

Table 8: **Stats.**

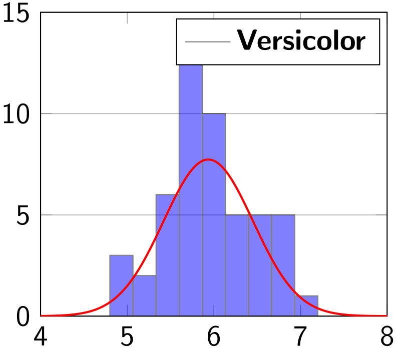



Figure 10: **Distribution**

# Problem state VII

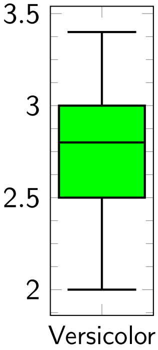



Figure 11: **Sepal width**

| Name | Value |
|---|---|
| **mean** | 2.77 |
| **std** | 0.31 |
| **max** | 3.40 |
| 75**%** | 3.00 |
| 50**%** | 2.80 |
| 25**%** | 2.53 |
| **min** | 2.00 |

Table 9: **Stats.**

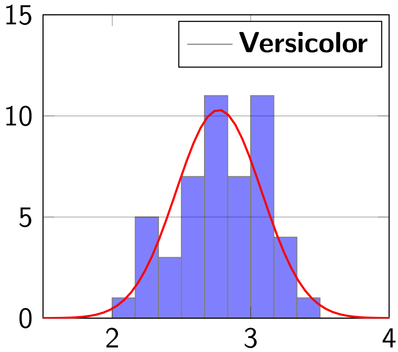



Figure 12: **Distribution**

# Problem state VIII

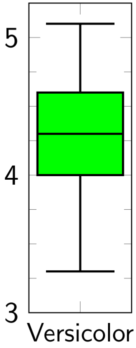



Figure 13: **Petal length**

| Name | Value |
|---|---|
| **mean** | 4.26 |
| **std** | 0.47 |
| **max** | 5.10 |
| **75%** | 4.60 |
| **50%** | 4.35 |
| **25%** | 4.00 |
| **min** | 3.00 |

Table 10: **Stats.**

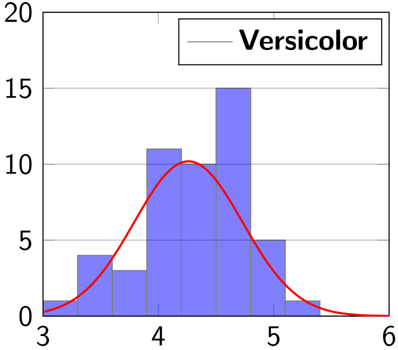



Figure 14: **Distribution**

# Problem state IX

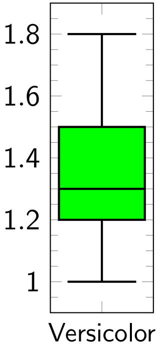



Figure 15: **Petal width**

| Name | Value |
|---|---|
| **mean** | 1.33 |
| **std** | 0.20 |
| **max** | 1.80 |
| 75**%** | 1.50 |
| 50**%** | 1.30 |
| 25**%** | 1.20 |
| **min** | 1.00 |

Table 11: **Stats.**

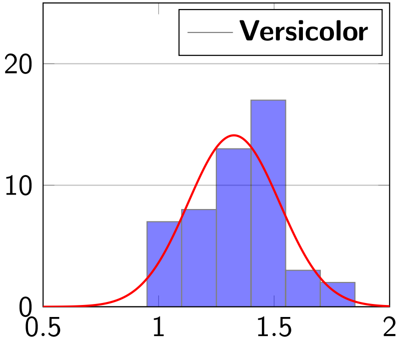



Figure 16: **Distribution**

# Problem state X

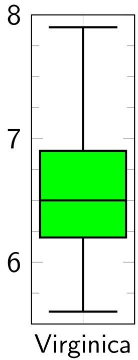



Figure 17: **Sepal length**

| Name | Value |
|---|---|
| **mean** | 6.59 |
| **std** | 0.64 |
| **max** | 7.90 |
| 75**%** | 6.90 |
| 50**%** | 6.50 |
| 25**%** | 6.23 |
| **min** | 4.90 |

Table 12: **Stats.**

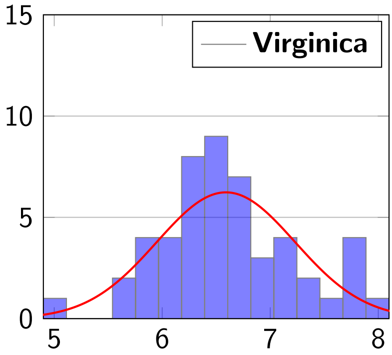



Figure 18: **Distribution**

# Problem state XI

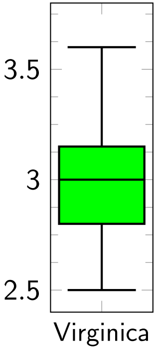



Figure 19: **Sepal width**

| **Name** | **Value** |
|---|---|
| **mean** | 2.97 |
| **std** | 0.32 |
| **max** | 3.80 |
| 75**%** | 3.18 |
| 50**%** | 3.00 |
| 25**%** | 2.80 |
| **min** | 2.20 |

Table 13: **Stats.**

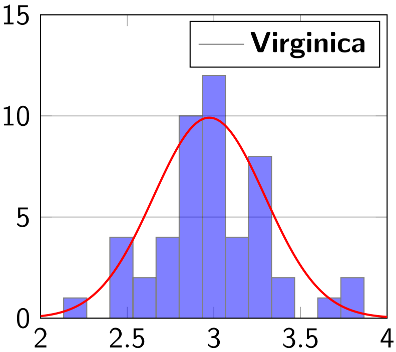



Figure 20: **Distribution**

# Problem state XII

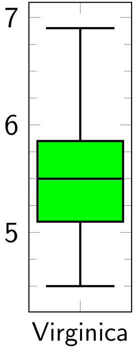



Figure 21: **Petal length**

| **Name** | **Value** |
|---|---|
| **mean** | 5.55 |
| **std** | 0.55 |
| **max** | 6.90 |
| 75**%** | 5.88 |
| 50**%** | 5.55 |
| 25**%** | 5.10 |
| **min** | 4.50 |

Table 14: **Stats.**

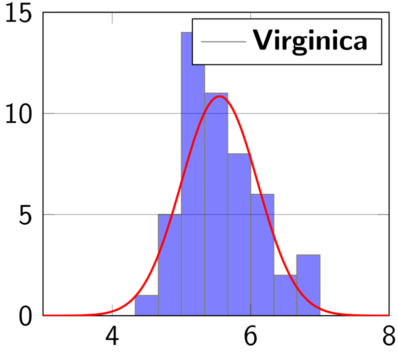



Figure 22: **Distribution**

# Problem state XIII

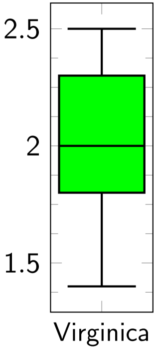



Figure 23: **Petal width**

| Name | Value |
|---|---|
| **mean** | 2.03 |
| **std** | 0.27 |
| **max** | 2.50 |
| 75**%** | 2.30 |
| 50**%** | 2.00 |
| 25**%** | 1.80 |
| **min** | 1.40 |

Table 15: **Stats.**

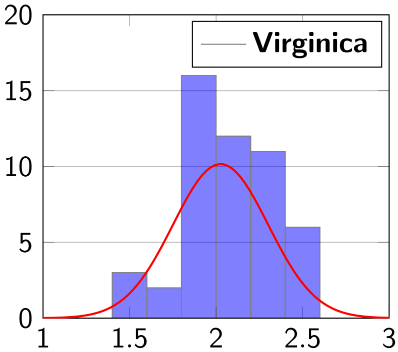
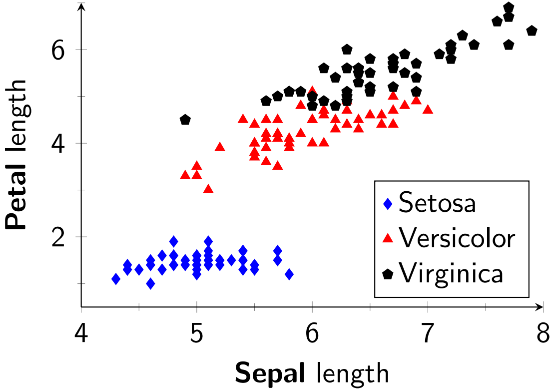
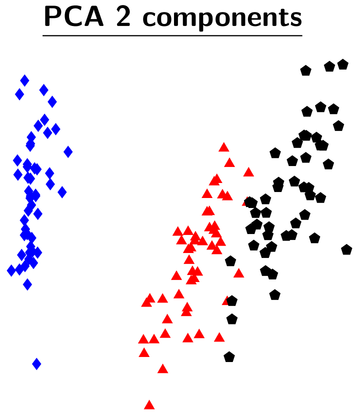



Figure 24: **Distribution**

# Scatter plot (left) and PCA (right) on iris dataset



# Shapiro – Wilk's normality test ([Razali et al., 2011])

**Prerequisites:**

**H0** normal.

**H1** not normal.

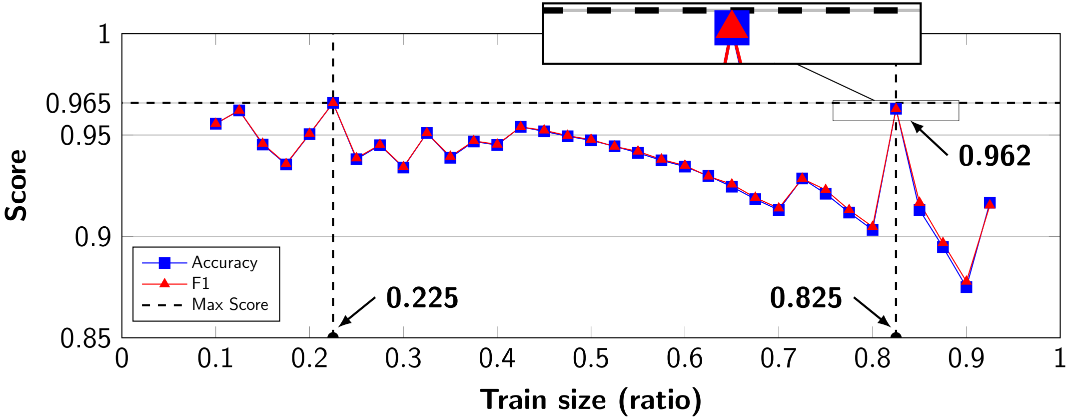
$\alpha$ 10%.

| | **Setosa** | **Versicolor** | **Virginica** |
|---|---|---|---|
| **Sepal len.** | 0.459 | 0.465 | 0.258 |
| **Sepal wid.** | 0.272 | 0.338 | 0.181 |
| **Petal len.** | 0.055 | 0.158 | 0.110 |
| **Petal wid.** | 0.000 | 0.027 | 0.087 |

Table 16: P-val for SW test statistics

**Conclusion:** under other things being equal data is normal.

# Fitting results



# Summary: topics covered

1. **Purposes** and **applications** of Bayesian Classifier.
2. **Maths** behind the algorithm.
3. **Limitations**.
4. **Example** on iris dataset.
5. **Best** score (**0.965** accuracy) reached at **0.225** train ratio.

# References

[Bayes and Price, 1763] Bayes, T. and Price, R. (1763).
An essay towards solving a problem in the doctrine of chances.
*Philosophical Transactions (1683-1775)*, 53:370–418.

[Razali et al., 2011] Razali, N. M., Wah, Y. B., et al. (2011).
Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests.
*Journal of statistical modeling and analytics*, 2(1):21–33.

[Wilks, 1938] Wilks, S. S. (1938).
The large-sample distribution of the likelihood ratio for testing composite hypotheses.
*The annals of mathematical statistics*, 9(1):60–62.

# Conclusions

## Thank you for attention!

Questions are welcome.