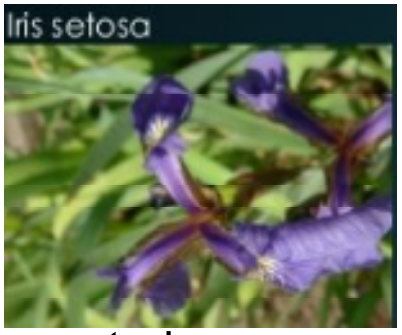
Review: 4-14, 4-19

- Types of machine learning problems
- Regularized Logistic Regression
- Naive Bayes Classifier
- Implementing a Gaussian Naives Bayes

- Application of probability, statistics, and prediction for measuring county mortality rates from Twitter.

Gaussian Naive Bayes



Assume $\mathrm{P}(\mathrm{X}|\mathrm{Y})$ is *Normal*

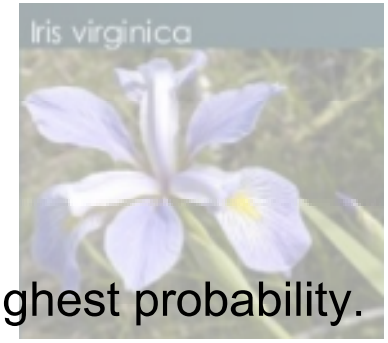
Then, training is:

1. Estimate $\mathrm{P}(\mathrm{Y} = \mathrm{k})$; $\boldsymbol{\pi}_{\mathrm{k}} = \mathrm{count}(\mathrm{Y} = \mathrm{k}) \,/\, \mathrm{Count}(\mathrm{Y} = *)$
2. MLE to find parameters ($\boldsymbol{\mu}$, $\boldsymbol{\sigma}$) for each class of Y.
   (the "class conditional distribution")

Maximum a Posteriori (MAP): Pick the class with the maximum posterior probability.

$$\hat{y} = arg\max_{y} \; \mathrm{P}(y) \prod_{i=1}^{m} \mathrm{P}(X_i|y)$$

Gaussian Naive Bayes



MLE: For which parameters does the observed data have the highest probability.

$$L(\theta) = \prod_{i=1}^{n} f(X_i; \theta) \qquad l(\theta) = log \sum_{i=1}^{n} f(X_i; \theta)$$

Maximum a Posteriori (MAP): Pick the class with the maximum posterior probability.

Unnormalized Posterior

$$\hat{y} = arg\max_{y} \mathrm{P}(y) \prod_{i=1}^{m} \mathrm{P}(X_i|y)$$

Gaussian Naive Bayes

Assume $P(X|Y)$ is *Normal*

Then, training is:

1. Estimate $P(Y = k)$; $\pi_k = count(Y = k) / Count(Y = *)$
2. Estimate $P(X_i|Y)$ for each feature (this "class-conditional distribution")

Maximum a Posteriori (MAP): Pick the class with the maximum posterior probability.

Without knowing $P(X)$, can we turn this into the (normalized) posterior?

Unnormalized Posterior

$$\hat{y} = arg\max_y \left[P(y) \prod_{i=1}^{m} P(X_i|y) \right]$$

Gaussian Naive Bayes

Use the Law of Total Probability, for all i = 1 ... k, where $A_1 \ldots A_k$ **partition** Ω:

Maximum a Posteriori (MAP): Pick the class with the maximum posterior probability.

Without knowing $\mathrm{P}(X)$, can we turn this into the (normalized) posterior?

Unnormalized Posterior

$$\hat{y} = arg\max_y \mathrm{P}(y) \prod_{i=1}^{m} \mathrm{P}(X_i|y)$$

Gaussian Naive Bayes

Use the Law of Total Probability, for all i = 1 ... k, where $A_1 \ldots A_k$ **partition** Ω:

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B, A_i)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\sum_{i=1}^{k} \mathrm{P}(B|A_i)\mathrm{P}(A_i)}$$

Maximum a Posteriori (MAP): Pick the class with the maximum posterior probability.

Without knowing $\mathrm{P}(X)$, can we turn this into the (normalized) posterior?

Unnormalized Posterior

$$\hat{y} = arg\max_{y} \mathrm{P}(y) \prod_{i=1}^{m} \mathrm{P}(X_i|y)$$

Gaussian Naive Bayesian Inference

Use the Law of Total Probability, for all i = 1 ... k, where $A_1 \ldots A_k$ **partition** Ω:

$$\mathrm{P}(A_i|B) = \frac{\mathrm{P}(B, A_i)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|A_i)\mathrm{P}(A_i)}{\sum_{i=1}^{k} \mathrm{P}(B|A_i)\mathrm{P}(A_i)}$$

discrete

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\mathrm{P}(A_i)}{\int \mathrm{P}(B|A)\mathrm{P}(A)dA}$$

continuous
A is "marginalized" out

Without knowing $\mathrm{P}(X)$, can we turn this into the (normalized) posterior?

Unnormalized Posterior

$$\hat{y} = arg\max_y \mathrm{P}(y) \prod_{i=1}^{m} \mathrm{P}(X_i|y)$$

Gaussian Naive Bayesian Inference

Q: What distinguishes Bayesian inference? **A**: Assume a

$\mathrm{P}(\theta)$ – prior

Bayesian Inference

$$Z = X_{training}$$

Given:

$\mathrm{P}(Z|\theta)$ – probability density or mass function (likelihood)
$\mathrm{P}(\theta)$ – prior

Goal: Compute the $\text{posterior} = \dfrac{(\text{prior})(\text{likelihood})}{\text{evidence}} = \dfrac{P(\theta)P(Z|\theta)}{P(Z)}$

Bayesian Inference

$$Z = X_{training}$$

Given:

$\mathrm{P}(Z|\theta)$ – probability density or mass function (likelihood)
$\mathrm{P}(\theta)$ – prior

Goal: Compute the $\text{posterior} = \frac{(\text{prior})(\text{likelihood})}{\text{evidence}} = \frac{P(\theta)P(Z|\theta)}{P(Z)}$

Types of priors:

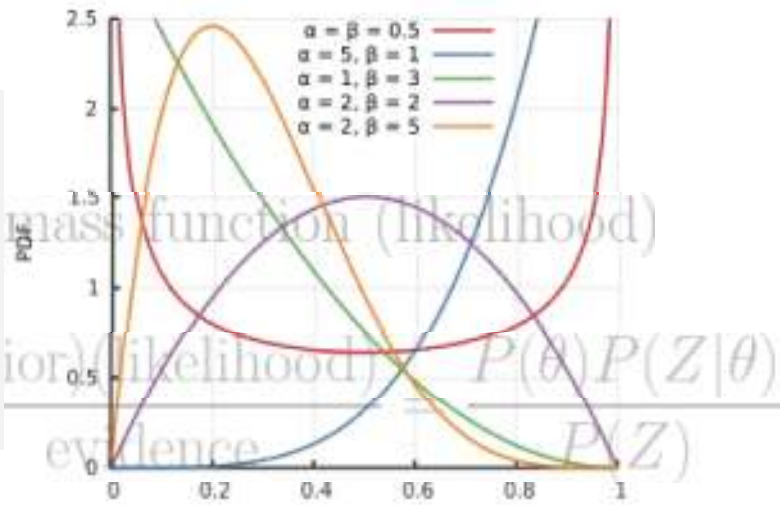
- Uninformative (Improper: not a probability (e.g. constant))
- Belief-based

- **Conjugate** to a likelihood: if the posterior is in the same family as the prior.

Bayesian Inference

Example: Beta(α, β) is conjugate to a Bernoulli likelihood.

https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions



Types of priors:

- Uninformative (Improper: not a probability (e.g. constant))
- Belief-based

- **Conjugate** to a likelihood: if the posterior is in the same family as the prior.

Bayesian Inference

$$Z = X_{training}$$

Given:

$\mathrm{P}(Z|\theta)$ – probability density or mass function (likelihood)
$\mathrm{P}(\theta)$ – prior

Goal: Compute the $\text{posterior} = \dfrac{(\text{prior})(\text{likelihood})}{\text{evidence}} = \dfrac{P(\theta)P(Z|\theta)}{P(Z)}$

Bayesian Inference

$$Z = X_{training}$$

Given:

$\mathrm{P}(Z|\theta)$ – probability density or mass function (likelihood)
$\mathrm{P}(\theta)$ – prior

Goal: Compute the $\text{posterior} = \dfrac{(\text{prior})(\text{likelihood})}{\text{evidence}} = \dfrac{P(\theta)P(Z|\theta)}{P(Z)}$

$$\mathrm{P}(\theta|Z) = \frac{\mathrm{P}(\theta)\mathrm{P}(Z|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(Z|\theta)d\theta}$$

Bayesian Inference

$$Z = X_{training}$$

Given:

$\mathrm{P}(Z|\theta)$ – probability density or mass function (likelihood)

$\mathrm{P}(\theta)$ – prior

Goal: Compute the $\text{posterior} = \dfrac{(\text{prior})(\text{likelihood})}{\text{evidence}} = \dfrac{P(\theta)P(Z|\theta)}{P(Z)}$

$$\mathrm{P}(\theta|Z) = \frac{\mathrm{P}(\theta)\mathrm{P}(Z|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(Z|\theta)d\theta}$$

$$\mathrm{P}(z^{new}|Z) = \int \mathrm{P}(z^{new}|\theta)\mathrm{P}(\theta|Z)d\theta$$ -- predictive distribution

Bayesian Inference

$$Z = X_{training}$$

Given:

$\mathrm{P}(Z|\theta)$ – probability density or mass function (likelihood)

$\mathrm{P}(\theta)$ – prior

Goal: Compute the $\text{posterior} = \dfrac{(\text{prior})(\text{likelihood})}{\text{evidence}} = \dfrac{P(\theta)P(Z|\theta)}{P(Z)}$

$$\mathrm{P}(\theta|Z) = \frac{\mathrm{P}(\theta)\mathrm{P}(Z|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(Z|\theta)d\theta}$$

Like a posterior-weighted average of $\mathrm{P}(Z^{new}|\theta)$

$$\mathrm{P}(z^{new}|Z) = \int \mathrm{P}(z^{new}|\theta)\mathrm{P}(\theta|Z)d\theta$$ -- predictive distribution