# Naïve Bayes
# Lecture 17

David Sontag
New York University

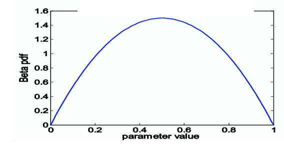
Slides adapted from Luke Zettlemoyer, Carlos Guestrin, Dan Klein, and Mehryar Mohri

# Bayesian Learning

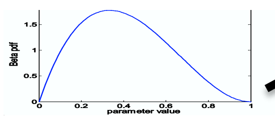
- Use Bayes' rule!

Data Likelihood

Prior




Posterior




$$P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

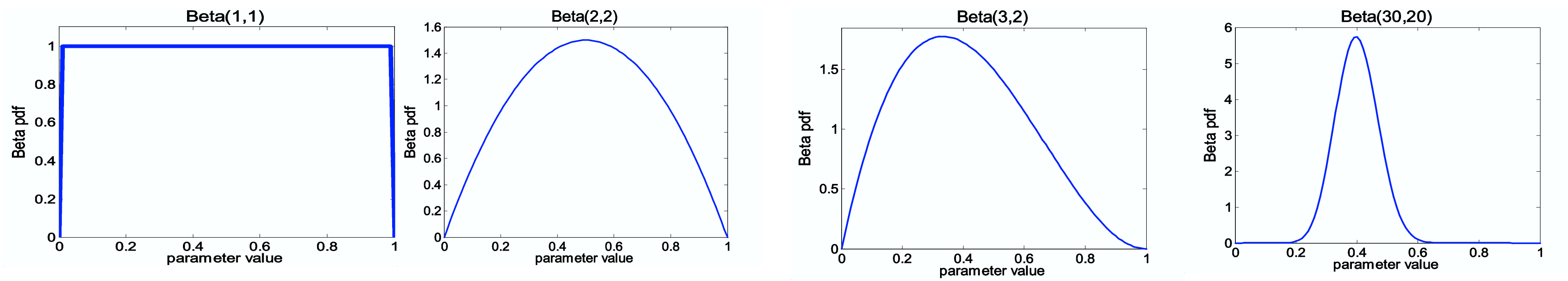
Normalization

- Or equivalently: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$
- For *uniform* priors, this reduces to maximum likelihood estimation!

$$P(\theta) \propto 1 \qquad P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)$$

# Prior + Data -> Posterior

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$




- Likelihood function: $P(\mathcal{D} \mid \theta) = \theta^{\alpha_H}(1-\theta)^{\alpha_T}$
- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta)$

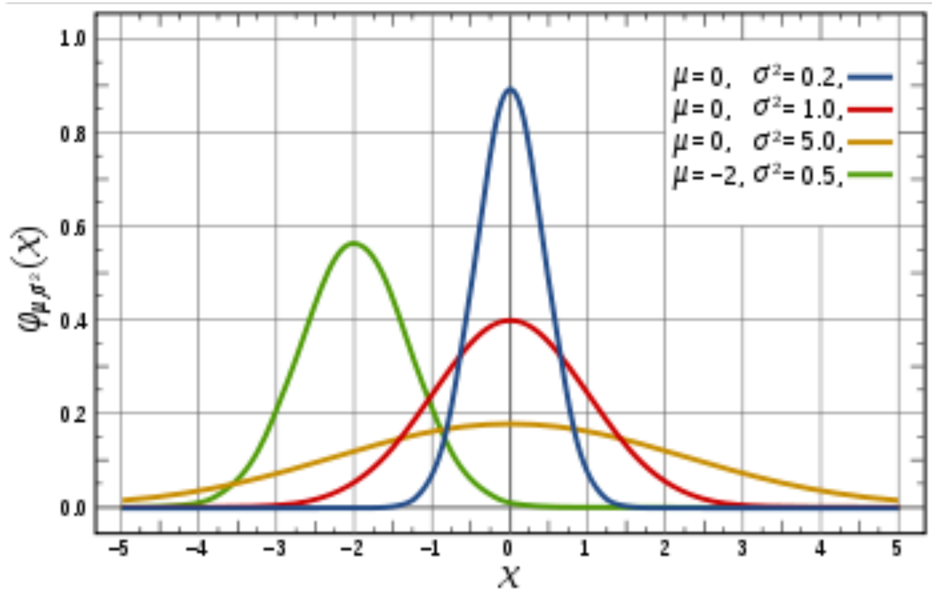
$$P(\theta \mid \mathcal{D}) \propto \theta^{\alpha_H}(1-\theta)^{\alpha_T}\ \theta^{\beta_H - 1}(1-\theta)^{\beta_T - 1}$$

$$= \theta^{\alpha_H + \beta_H - 1}(1-\theta)^{\alpha_T + \beta_t + 1}$$

$$= Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

# What about continuous variables?

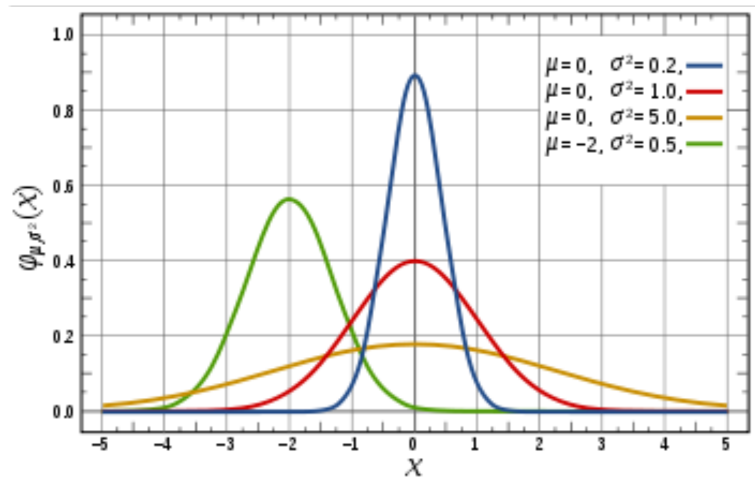
- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- **You say: Let me tell you about Gaussians…**




$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

# Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) are Gaussian
  - $X \sim N(\mu,\sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu+b, a^2\sigma^2)$




- Sum of Gaussians is Gaussian
  - $X \sim N(\mu_X,\sigma^2_X)$
  - $Y \sim N(\mu_Y,\sigma^2_Y)$
  - $Z = X+Y \rightarrow Z \sim N(\mu_X+\mu_Y, \sigma^2_X+\sigma^2_Y)$

- Easy to differentiate, as we will see soon!

# Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores
- Learn parameters
  - Mean: $\mu$
  - Variance: $\sigma$

| $x_i$ $i =$ | Exam Score |
|---|---|
| 0 | 85 |
| 1 | 95 |
| 2 | 100 |
| 3 | 12 |
| … | … |
| 99 | 89 |

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

# MLE for Gaussian: $P(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

- Prob. of i.i.d. samples $D=\{x_1,\ldots,x_N\}$:

$$P(\mathcal{D} \mid \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\mu_{MLE}, \sigma_{MLE} = \arg\max_{\mu,\sigma} P(\mathcal{D} \mid \mu, \sigma)$$

- Log-likelihood of data:

$$\begin{aligned} \ln P(\mathcal{D} \mid \mu, \sigma) &= \ln \left[ \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}} \right] \\ &= -N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

# Your second learning algorithm: MLE for mean of a Gaussian

- What's MLE for mean?

$$\frac{d}{d\mu}\ln P(\mathcal{D}\mid\mu,\sigma) = \frac{d}{d\mu}\left[-N\ln\sigma\sqrt{2\pi} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{d}{d\mu}\left[-N\ln\sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N}\frac{d}{d\mu}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^2} = 0$$

$$= -\sum_{i=1}^{N}x_i + N\mu = 0$$

$$\boxed{\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}x_i}$$

# MLE for variance

- Again, set derivative to zero:

$$\frac{d}{d\sigma}\ln P(\mathcal{D} \mid \mu,\sigma) = \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi} - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= \frac{d}{d\sigma}\left[-N \ln \sigma\sqrt{2\pi}\right] - \sum_{i=1}^{N}\frac{d}{d\sigma}\left[\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$= -\frac{N}{\sigma} + \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{\sigma^3} \qquad = 0$$

$$\boxed{\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\widehat{\mu})^2}$$

# Learning Gaussian parameters

- MLE:

$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

$$\widehat{\sigma}^2_{unbiased} = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

# Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

$$P(\mu \mid \eta, \lambda) = \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}$$

# Bayesian Prediction

- **Definition**: the expected conditional loss of predicting $\widehat{y} \in \mathcal{Y}$ is

$$\mathcal{L}[\widehat{y}|\mathbf{x}] = \sum_{y \in \mathcal{Y}} L(\widehat{y}, y) \Pr[y|\mathbf{x}].$$

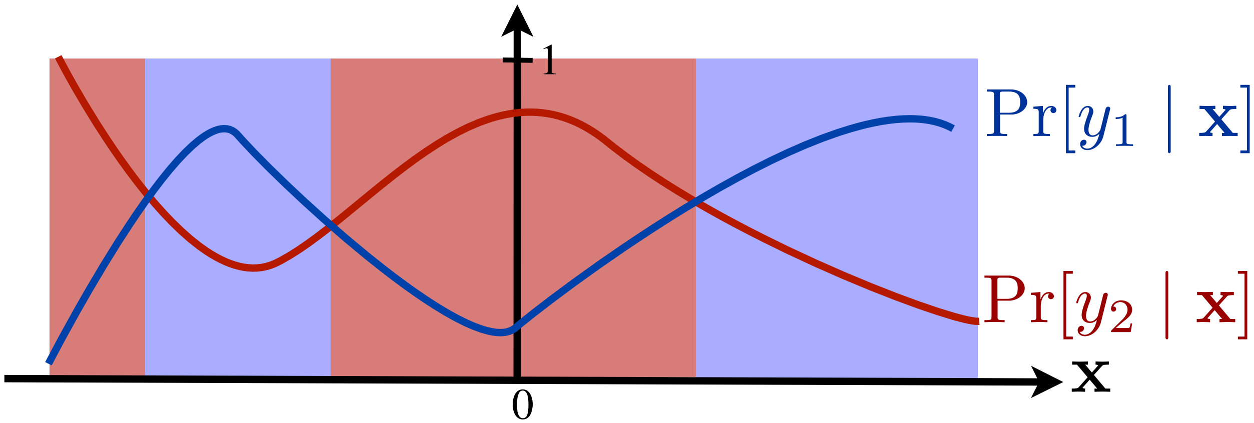
- **Bayesian decision**: predict class minimizing expected conditional loss, that is

$$\widehat{y}^* = \operatorname*{argmin}_{\widehat{y}} \mathcal{L}[\widehat{y}|\mathbf{x}] = \operatorname*{argmin}_{\widehat{y}} \sum_{y \in \mathcal{Y}} L(\widehat{y}, y) \Pr[y|\mathbf{x}].$$

  - zero-one loss: $\widehat{y}^* = \operatorname*{argmax}_{\widehat{y}} \Pr[\widehat{y}|\mathbf{x}].$

$\longrightarrow$ Maximum a Posteriori (MAP) principle.

# Binary Classification - Illustration



# Maximum a Posteriori (MAP)

- Definition: the MAP principle consists of predicting according to the rule
$$\widehat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} \Pr[y|\mathbf{x}].$$
- Equivalently, by the Bayes formula:
$$\widehat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} \frac{\Pr[\mathbf{x}|y]\Pr[y]}{\Pr[\mathbf{x}]} = \boxed{\operatorname*{argmax}_{y \in \mathcal{Y}} \Pr[\mathbf{x}|y]\Pr[y]}.$$

→ How do we determine $\Pr[\mathbf{x}|y]$ and $\Pr[y]$ ?
Density estimation problem.

# Density Estimation

- **Data**: sample drawn i.i.d. from set $X$ according to some distribution $D$,

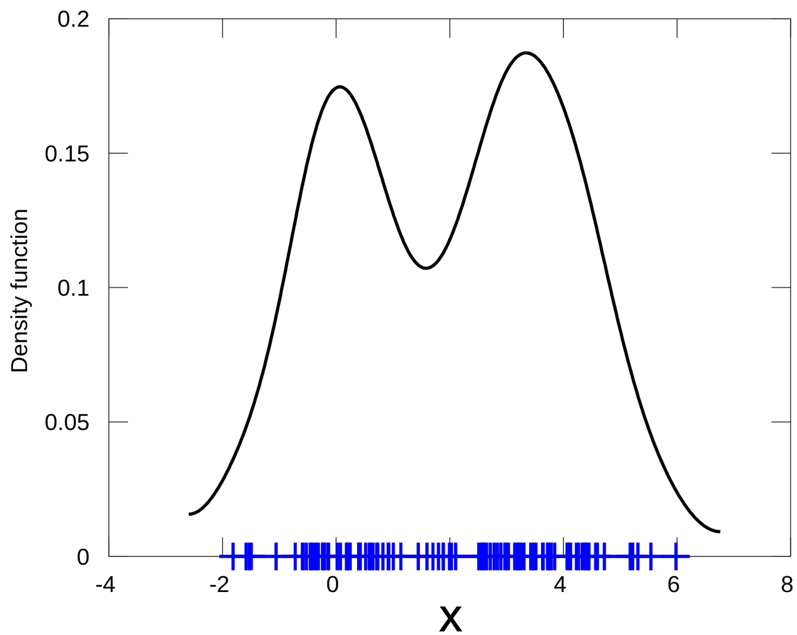
$$x_1, \ldots, x_m \in X.$$

- **Problem**: find distribution $p$ out of a set $\mathcal{P}$ that best estimates $D$.

[Slide from Mehryar Mohri]

# Density estimation

- Can make **parametric** assumption, e.g. that Pr(**x**|y) is a multivariate Gaussian distribution
- When the dimension of **x** is small enough, can use a **non-parametric** approach (e.g., *kernel density estimation*)



# Difficulty of (naively) estimating high-dimensional distributions

- Can we directly estimate the data distribution P(X,Y)?
- How do we represent these? How many parameters?
  - Prior, P(Y):
    - Suppose Y is composed of *k* classes
  - Likelihood, P(**X**|Y):
    - Suppose **X** is composed of *n* binary features

| mpg | cylinders | displacement | horsepower | weight | acceleration | modelyear | maker |
|---|---|---|---|---|---|---|---|
| | | | | | | | |
| good | 4 | low | low | low | high | 75to78 | asia |
| bad | 6 | medium | medium | medium | medium | 70to74 | america |
| bad | 4 | medium | medium | medium | low | 75to78 | europe |
| bad | 8 | high | high | high | low | 70to74 | america |
| bad | 6 | medium | medium | medium | medium | 70to74 | america |
| bad | 4 | low | medium | low | medium | 70to74 | asia |
| bad | 4 | low | medium | low | low | 70to74 | asia |
| bad | 8 | high | high | high | low | 75to78 | america |
| : | : | : | : | : | : | : | : |
| : | : | : | : | : | : | : | : |
| : | : | : | : | : | : | : | : |
| bad | 8 | high | high | high | low | 70to74 | america |
| good | 8 | high | medium | high | high | 79to83 | america |
| bad | 8 | high | high | high | low | 75to78 | america |
| good | 4 | low | low | low | low | 79to83 | america |
| bad | 6 | medium | medium | medium | high | 75to78 | america |
| good | 4 | medium | low | low | low | 79to83 | america |
| good | 4 | low | low | medium | high | 79to83 | america |
| bad | 8 | high | high | high | low | 70to74 | america |
| good | 4 | low | medium | low | medium | 75to78 | europe |
| bad | 5 | medium | medium | medium | medium | 75to78 | europe |

- Complex model $\rightarrow$ High variance with limited data!!!

# Conditional Independence

- X is **conditionally independent** of Y given Z, if the probability distribution for X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = i | Y = j, Z = k) = P(X = i | Z = k)$$

- e.g.,

$$P(Thunder | Rain, Lightning) = P(Thunder | Lightning)$$

- Equivalent to:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

# Naïve Bayes

- Naïve Bayes assumption:
  - Features are independent given class:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

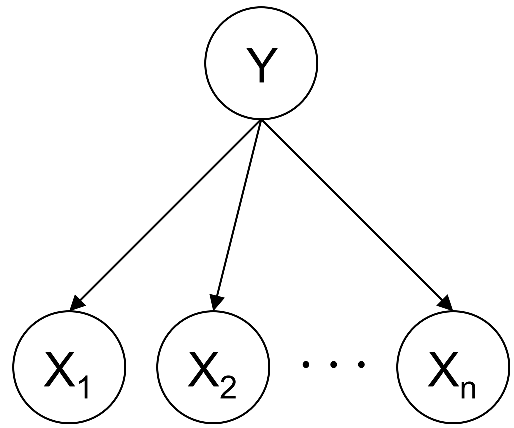
  - More generally:

$$\boxed{P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)}$$

- How many parameters now?
  - Suppose **X** is composed of *n* binary features

# The Naïve Bayes Classifier

- Given:
  - Prior P(Y)
  - *n* conditionally independent features **X** given the class Y
  - For each $X_i$, we have likelihood $P(X_i|Y)$




- Decision rule:

$$
\begin{aligned}
y^* = h_{NB}(\mathbf{x}) &= \arg\max_y P(y) P(x_1, \ldots, x_n \mid y) \\
&= \arg\max_y P(y) \prod_i P(x_i | y)
\end{aligned}
$$

If certain assumption holds, NB is optimal classifier!
(they typically don't)