Introduction to Machine Learning
Lecture 3

Mehryar Mohri
Courant Institute and Google Research
mohri@cims.nyu.edu

Bayesian Learning

Bayes' Formula/Rule

- Terminology:

$$\underbrace{\Pr[Y \mid X]}_{\text{posterior probability}} = \frac{\overbrace{\Pr[X \mid Y]}^{\text{likelihood}}\,\overbrace{\Pr[Y]}^{\text{prior}}}{\underbrace{\Pr[X]}_{\text{evidence}}}.$$

Loss Function

- Definition: function $L \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ indicating the penalty for an incorrect prediction.
  - $L(\widehat{y}, y)$: loss for prediction of $\widehat{y}$ instead of y.
- Examples:
  - zero-one loss: standard loss function in classification; $L(y, y') = 1_{y \neq y'}$ for $y, y' \in \mathcal{Y}$.
  - non-symmetric losses: e.g., for spam classification; $L(\widehat{\text{ham}}, \text{spam}) \leq L(\widehat{\text{spam}}, \text{ham})$.
  - squared loss: standard loss function in regression; $L(y, y') = (y' - y)^2$.

Classification Problem

- Input space $\mathcal{X}$: e.g., set of documents.
  - feature vector $\Phi(x) \in \mathbb{R}^N$ associated to $x \in \mathcal{X}$.
  - notation: feature vector $\mathbf{x} \in \mathbb{R}^N$.
  - example: vector of word counts in document.
- Output or target space $\mathcal{Y}$: set of classes; e.g., `sport`, `business`, `art`.
- Problem: given $\mathbf{x}$, predict the correct class $y \in \mathcal{Y}$ associated to $\mathbf{x}$.

Bayesian Prediction

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

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

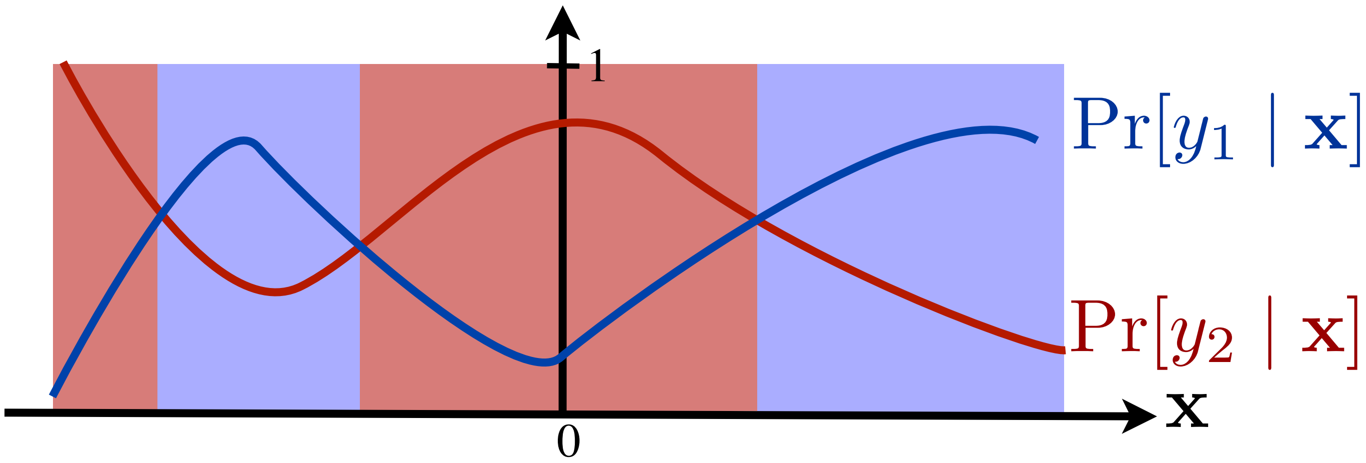
- **Bayesian decision**: predict class minimizing expected conditional loss, that is

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

  - 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.

Application - Maximum a Posteriori

- **Formulation**: hypothesis set H.

$$\hat{h} = \operatorname*{argmax}_{h \in H} \Pr[h|O] = \operatorname*{argmax}_{h \in H} \frac{\Pr[O|h]\Pr[h]}{\Pr[O]} = \operatorname*{argmax}_{h \in H} \Pr[O|h]\Pr[h].$$

- **Example**: determine if a patient has a rare disease $H = \{d, nd\}$, given laboratory test $O = \{pos, neg\}$. With $\Pr[d] = .005, \Pr[pos|d] = .98, \Pr[neg|nd] = .95$, if the test is positive, what should be the diagnosis?

$$\Pr[pos|d] \Pr[d] = .98 \times .005 = .0049.$$

$$\Pr[pos|nd] \Pr[nd] = (1 - .95) \times .(1 - .005) = .04975 > .0049.$$

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 .
  - Note: we will study density estimation specifically in a future lecture.

Maximum Likelihood

- Likelihood: probability of observing sample under distribution $p \in \mathcal{P}$, which, given the independence assumption is

$$\Pr[x_1, \ldots, x_m] = \prod_{i=1}^{m} p(x_i).$$

- Principle: select distribution maximizing sample probability

$$p_\star = \operatorname*{argmax}_{p \in \mathcal{P}} \prod_{i=1}^{m} p(x_i),$$

or $$p_\star = \operatorname*{argmax}_{p \in \mathcal{P}} \sum_{i=1}^{m} \log p(x_i).$$

Example: Bernoulli Trials

- Problem: find most likely Bernoulli distribution, given sequence of coin flips

$$H, T, T, H, T, H, T, H, H, H, T, T, \ldots, H.$$

- Bernoulli distribution: $p(H) = \theta, p(T) = 1 - \theta.$

- Likelihood: $$\begin{aligned} l(p) &= \log \theta^{N(H)} (1-\theta)^{N(T)} \\ &= N(H) \log \theta + N(T) \log(1-\theta). \end{aligned}$$

- Solution: l is differentiable and concave;

$$\frac{dl(p)}{d\theta} = \frac{N(H)}{\theta} - \frac{N(T)}{1-\theta} = 0 \Leftrightarrow \theta = \frac{N(H)}{N(H) + N(T)}.$$

Example: Gaussian Distribution

- **Problem:** find most likely Gaussian distribution, given sequence of real-valued observations

$$3.18, 2.35, .95, 1.175, \ldots$$

- **Normal distribution:** $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$
- **Likelihood:** $l(p) = -\frac{1}{2} m \log(2\pi\sigma^2) - \sum_{i=1}^{m} \frac{(x_i - \mu)^2}{2\sigma^2}.$
- **Solution:** l is differentiable and concave;

$$\frac{\partial p(x)}{\partial \mu} = 0 \Leftrightarrow \mu = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \frac{\partial p(x)}{\partial \sigma^2} = 0 \Leftrightarrow \sigma^2 = \frac{1}{m}\sum_{i=1}^{m} x_i^2 - \mu^2.$$

ML Properties

- Problems:
  - the underlying distribution may not be among those searched.
  - overfitting: number of examples too small wrt number of parameters.
  - $\Pr[y] = 0$ if class y does not appear in sample!

    $\longrightarrow$ smoothing techniques.

Additive Smoothing

- Definition: the additive or Laplace smoothing for estimating $\Pr[y]$, $y \in \mathcal{Y}$, from a sample of size m is defined by

$$\widehat{\Pr}[y] = \frac{|y| + \alpha}{m + \alpha|\mathcal{Y}|}.$$

  - $\alpha = 0$: ML estimator (MLE).
  - MLE after adding α to the count of each class.
  - Bayesian justification based on Dirichlet prior.
  - poor performance for some applications, such as *n*-gram language modeling.

Estimation Problem

- Conditional probability: $\Pr[\mathbf{x} \mid y] = \Pr[x_1, \ldots, x_N \mid y]$.
  - for large N, number of features, difficult to estimate.
  - even if features are Boolean, that is $x_i \in \{0, 1\}$, there are 2^N possible feature vectors!

  $\longrightarrow$ may need very large sample.

Naive Bayes

这里就是朴素的意思：假设变量相互独立

- Conditional independence assumption: for any $y \in \mathcal{Y}$,

$$\Pr[x_1, \ldots, x_N \mid y] = \Pr[x_1 \mid y] \ldots \Pr[x_N \mid y].$$

  - given the class, the features are assumed to be independent.
  - strong assumption, typically does not hold.

Example - Document Classification

- Features: presence/absence of word x_i.
- Estimation of $\Pr[x_i \mid y]$: frequency of word x_i among documents labeled with y, or smooth estimate.
- Estimation of $\Pr[y]$: frequency of class y in sample.
- Classification:

$$\widehat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} \Pr[y] \prod_{i=1}^{N} \Pr[x_i \mid y].$$

Naive Bayes - Binary Classification

- Classes: $\mathcal{Y} = \{-1, +1\}$.
- Decision based on sign of $\log \frac{\Pr[+1|\mathbf{x}]}{\Pr[-1|\mathbf{x}]}$; in terms of log-odd ratios:

$$\begin{aligned}
\log \frac{\Pr[+1 \mid \mathbf{x}]}{\Pr[-1 \mid \mathbf{x}]} &= \log \frac{\Pr[+1] \Pr[\mathbf{x} \mid +1]}{\Pr[-1] \Pr[\mathbf{x} \mid -1]} \\
&= \log \frac{\Pr[+1] \prod_{i=1}^{N} \Pr[x_i \mid +1]}{\Pr[-1] \prod_{i=1}^{N} \Pr[x_i \mid -1]} \\
&= \log \frac{\Pr[+1]}{\Pr[-1]} + \sum_{i=1}^{N} \underbrace{\log \frac{\Pr[x_i \mid +1]}{\Pr[x_i \mid -1]}}_{\text{contribution of feature/expert } i \text{ to decision}}.
\end{aligned}$$

contribution of feature/expert i to decision

Naive Bayes = Linear Classifier

- **Theorem**: assume that $x_i \in \{0, 1\}$ for all $i \in [1, N]$. Then, the Naive Bayes classifier is defined by

$$\mathbf{x} \mapsto \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x} + b),$$

where $w_i = \log \frac{\Pr[x_i=1|+1]}{\Pr[x_i=1|-1]} - \log \frac{\Pr[x_i=0|+1]}{\Pr[x_i=0|-1]}$

and $b = \log \frac{\Pr[+1]}{\Pr[-1]} + \sum_{i=1}^{N} \log \frac{\Pr[x_i=0|+1]}{\Pr[x_i=0|-1]}$.

- **Proof**: observe that for any $i \in [1, N]$,

$$\log \frac{\Pr[x_i \mid +1]}{\Pr[x_i \mid -1]} = \left(\log \frac{\Pr[x_i = 1 \mid +1]}{\Pr[x_i = 1 \mid -1]} - \log \frac{\Pr[x_i = 0 \mid +1]}{\Pr[x_i = 0 \mid -1]}\right) x_i + \log \frac{\Pr[x_i = 0 \mid +1]}{\Pr[x_i = 0 \mid -1]}.$$

Summary

- Bayesian prediction:
  - requires solving density estimation problems.
  - often difficult to estimate $\Pr[\mathbf{x} \mid y]$ for $\mathbf{x} \in \mathbb{R}^N$.
  - but, simple and easy to apply; widely used.
- Naive Bayes:
  - strong assumption.
  - straightforward estimation problem.
  - specific linear classifier.
  - sometimes surprisingly good performance.