Introduction to Machine Learning
236756
Nir Ailon

Lecture 11:
Probabilistic Models
Most of the Course So Far:
Discriminative Approach

• (Unknown) distribution $D$ over $\mathcal{X} \times \mathcal{Y}$
• We only cared about $\Pr[Y = y|X = x]$
  • Shorthand notation: $\Pr[Y|X]$
• This allows us to predict
  • E.g. in binary case, given $x \in \mathcal{X}$

$$h(x) = \begin{cases} 
1 & \Pr[Y = 1|X = x] > \frac{1}{2} \\
0 & \text{otherwise}
\end{cases}$$

“Bayes Optimal”
ERM Can Sometimes Be Viewed as Discriminative Approach for a "Made Up" Probabilistic Method

• Linear regression alternative interpretation:
  • \( (Y|X = x) \sim N(w \cdot x, 1) \)
  • Choose \( w \) minimizing \(-\log \prod_{i=1}^{m} f_w(y_i|x_i)\)
  • \( f_w(y|x) = \frac{1}{\sqrt{2\pi}} e^{- (y - w \cdot x)^2 / 2} \)

• Logistic regression alternative interpretation:
  • \( Y \) binary
  • \( (Y|X = x) \sim \text{sgn} (\text{logistic}(w \cdot x, \gamma)) \)
  • Choose \( w \) minimizing \(-\text{LogLikelihood}...\)

• Some questions:
  • Does hinge loss have a probabilistic interpretation?
  • What about regularization?
Should We Care About $\Pr[X]$?

- Assume $\mathcal{X}$ discrete
- Applying Bayes rule:
  \[
  \Pr[Y = y | X = x] = \frac{\Pr[Y = y \land X = x]}{\Pr[X = x]}
  \]
  
  More simply:
  \[
  \Pr[Y = y | X = x] \propto \Pr[X = x | Y = y] \Pr[Y = y]
  \]

- Generative approach: Estimate $\Pr[X|Y], \Pr[Y]$
  .......implies estimating $\Pr[Y|X]$
Why Not Generative Approach

• Assume $\mathcal{X} = \{0,1\}^n$, $\mathcal{Y} = \{0,1\}$

• In general need exponentially many parameters to learn $P[X|Y]$

• “When solving a given problem, try to avoid a more general problem as an intermediate step”

/Vladimir Vapnik
Why Generative Approach?

• Distribution might come from parametric family. Notation:
  \( \Pr[X, Y|\Theta], \Pr[X|Y, \Theta], \Pr_\Theta[X, Y], \Pr_\Theta[X|Y] \)
  • Parallel with course so far:
    Think of parameter space as concept class

• Might be easy to learn parameters
  • Sample complexity
  • Computational complexity

• Learning process that generates \( X \) might help us for future prediction problems on same data
Stats 101: Maximum Likelihood Estimator (MLE)

• Given independent draws $x_1, \ldots, x_m$ of distribution from family with parameter $\Theta$

• Likelihood of $x_1, \ldots, x_m$ is function of $\Theta$ defined as
  \[ L(x_1 \ldots x_m; \Theta) = Pr_\Theta[x_1 \ldots x_m] \]

• Maximum likelihood estimator defined as
  \[ \hat{\Theta} = \arg\max_{\Theta} L(x_1 \ldots x_m; \Theta) \]

• Equivalently (and usually more conveniently):
  \[ \hat{\Theta} = \arg\max_{\Theta} \log L(x_1 \ldots x_m; \Theta) \]
Example: MLE For Biased Coin

- Coin with $\Theta = \text{Pr}[\text{HEADS}]$
- Given $x_1 \ldots x_m$ what is MLE $\hat{\Theta}$?
- Answer:
  $$\hat{\Theta} = \frac{1}{m} \#\{i : x_i = \text{HEADS}\}$$
- Proof:
  $$\log \text{Pr}_\Theta[x_1 \ldots x_m] = \#\text{HEADS} \cdot \log \Theta + \#\text{TAILS} \cdot \log(1 - \Theta)$$
- Derive wrt $\Theta$ and equate with 0

- Note: Can easily "overfit"
  If $\Theta$ small then w.p. $(1 - \Theta)^m$ will see all 0’s and get $-\infty$ in expression. Will get back to this later...
MLE for Continuous R.V.’s

- Replace probability with density
- E.g. Gaussian with θ = (μ, σ)

\[ Pr_θ[X = x] = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2} \]

\[ L(x_1 \ldots x_m; θ) = -\frac{1}{\sigma \sqrt{2\pi}} \sum (x_i - μ)^2 - m \cdot \log \sigma \sqrt{2\pi} \]

- MLE \( \hat{θ} = (\hat{μ}, \hat{σ}) \) given by

\[ \hat{μ} = \frac{1}{m} \sum x_i \quad \hat{σ} = \sqrt{\frac{1}{m} \sum (x_i - \hat{μ})^2} \]
Naïve Bayes Approach

• Assume $X = \{0,1\}^n, Y = \{0,1\}$
• Remember from Bayes rule:
  \[ \Pr[y|x] \propto \Pr[x|y] \Pr[y] = \Pr[x, y] \]
• Can’t learn $\Pr[x|y]$ for all $(x, y) \in \{0,1\}^{n+1}$
• Let’s assume

\[
\Pr[X = x|Y = y] = \prod_{j=1}^{n} \Pr[X[j] = x[j]|Y = y]
\]

\[ \Rightarrow h(x) = \arg\max_{y\in\{0,1\}} \log \Pr[y|x] \]
\[ = \arg\max_{y} \log \Pr[x|y] \Pr[y] \]
\[ = \arg\max_{y} \log P[y] + \sum \Pr[x[j]|y] \]

• Number of parameters = $2n + 1$ only!
Naïve Bayes Classifier
(Binary Case)

• Using "biased coin" probability estimation:
  • For each $j \in [n]$ estimate
    • $\theta_0[j] = \Pr[X[j] = 1|Y = 0]$
    • $\theta_1[j] = \Pr[X[j] = 1|Y = 1]$
  • Estimate $\Theta = \Pr[Y = 1]$

• Using last slide: We make prediction $h(x) = 1$ iff
  • $\log \Pr[1|x] > \log \Pr[0|x]$
  $\iff \log \Pr[x, 1] - \log \Pr[x, 0] > 0$
  $\iff \log \hat{\Theta} + \sum (x[j] \log \hat{\Theta}_1[j] + (1 - x[j]) \log (1 - \hat{\Theta}_1[j]))$
    - $\log (1 - \hat{\Theta}) - \sum (x[j] \log \hat{\Theta}_0[j] + (1 - x[j]) \log (1 - \hat{\Theta}_0[j])) > 0$

It’s a linear model!
Naïve Bayes Classifier (Gaussian Case)

- $X$ continuous
- For simplicity assume $Y \in \{0,1\}$
- $(X[j] | Y = y) \sim N(\mu_{jy}, \sigma_j)$
  - Estimate the $\mu$’s and the $\sigma$’s from the data
- $\forall j \neq j'; X[j], X[j']$ independent conditioned on $Y$ (naïveté)

$h(x) = \arg\max_y \log \Pr[y|x]$  
$= \arg\max_y \log \Pr[y] + \log \Pr[x|y]$  
$= \arg\max_y \log \Pr[y] + \log \Pr[x|y] - \log \Pr[y] - \log \Pr[x|\bar{y}]$  
$= \arg\max_y \log \frac{\Pr[y]}{\Pr[\bar{y}]} + \log \frac{\Pr[x|y]}{\Pr[x|\bar{y}]}$  
$= \arg\max_y \log \frac{\Pr[y]}{\Pr[\bar{y}]} + \sum_{j=1}^m \left( -\frac{(x_j - \mu_{jy})^2}{\sigma_j^2} + \frac{(x_j - \mu_{j\bar{y}})^2}{\sigma_j^2} \right)$  
$= \arg\max_y c + \sum_{j=1}^m 2(x - \mu_{jy})x_j / \sigma_j^2$

It’s a linear model!
(Gaussian) Naïve Bayes vs Linear Regression

• Both models predict using a linear model
• Both models choose linear coefficients by minimizing squared loss
  • Naïve Bayes fits $X|Y$ (assuming conditional indep.)
    • Generative!
  • Linear Regression fits $Y|X$
    • Discriminative!
Linear Discriminant Analysis (LDA)

• $X \in \mathbb{R}^n$

• If $X|Y$ assumed Gaussian, we saw that conditional independence of the coordinates gives rise to $O(n)$ parameters

• More generally, a vector Gaussian has $O(n^2)$ parameters
  • The expectation $\mu = (\mu[1], \ldots, \mu[n]) \in \mathbb{R}^n$
  • The covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$
    • Positive-(Semi)Definite

• Estimating $\mu, \Sigma$ from data can be done similarly for 1d Gaussians

• Conditional independence = Diagonal $\Sigma$
LDA

- Assume same covariance $\Sigma$ conditioned on all $y$
- Different expectation vectors $\mu_y$
- $\Pr[X = x|Y = y] = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^\frac{1}{2}} \exp\left\{-\frac{1}{2} (x - \mu_y)^T \Sigma^{-1} (x - \mu_y) \right\}$
- Will predict $h(x) = 1$ iff
  $$\log \left( \frac{\Pr[Y = 1] \Pr[x|Y = 1]}{\Pr[Y = 0] \Pr[x|Y = 0]} \right) > 0$$
- (Assuming $\Pr[Y = 1] = \Pr[Y = 0] = \frac{1}{2}$), iff:
  $$\frac{1}{2} (x - \mu_0)^T \Sigma^{-1} (x - \mu_0) - \frac{1}{2} (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) > 0$$
- Iff $\langle w, x \rangle + b > 0$, where
  $$w = (\mu_1 - \mu_0)^T \Sigma^{-1} \quad b = \frac{1}{2} (\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1)$$

Assuming binary $Y$ and $\Pr[Y = 1] = \Pr[Y = 0] = \frac{1}{2}$, MLE estimation: LDA equivalent to linear regression

It's a linear model!
Bayesian Reasoning

• Until now we thought of the distribution parameter(s) $\Theta$ as an unknown that we want to estimate optimally using MLE

• Reminder: This approach easily overfits
  • If you saw 3 HEADS and 0 TAILS, MLE ``thinks” TAILS should never appear (``The Black Swan” phenomenon)

• How do we fix this?

• In Bayesian Reasoning we think of $\Theta$ itself as a r.v.

• $\Pr[\Theta]$: “Prior Distribution”
  • Chosen by learner before learning
  • Samples $x_1 \ldots x_m$ independent only conditionally on $\Theta$
Bayesian Priors vs SRM

• Imposing a prior can be used to assign more weight to less complex models, less weight to complex models ⇒ Prefer simpler models

• Gaussian prior on $\Theta \in \mathbb{R}^n$
  • $-\text{LogLikelihood} \equiv L_2$ regularization

• Laplace prior on $\Theta \in \mathbb{R}^n$ ($f(x) \propto e^{-|x|}$)
  • $-\text{LogLikelihood} \equiv L_1$ regularization
Bayesian Reasoning

- \( \Pr[x|x_1 \ldots x_m] = \sum_\Theta \Pr[x|\Theta, x_1 \ldots x_m] \Pr[\Theta|x_1 \ldots x_m] \)
  
- \( = \sum_\Theta \Pr[x|\Theta] \Pr[\Theta|x_1 \ldots x_m] \)
  
- \( = \sum_\Theta \Pr[x|\Theta] \frac{\Pr[x_1 \ldots x_m|\Theta] \Pr[\Theta]}{\Pr[x_1 \ldots x_m]} \)
  
- \( = \sum_\Theta \Pr[x|\Theta] \frac{\Pr[\Theta] \prod \Pr[x_i|\Theta]}{\Pr[x_1 \ldots x_m]} \)
  
- \( \propto \sum_\Theta \Pr[x|\Theta] \Pr[\Theta] \prod \Pr[x_i|\Theta] \)

- Example:
  
  - \( \Pr[HEADS|\Theta] = \Theta, \ \Theta \sim Unif([0,1]) \)
  
  - \( \Pr[HEADS|x_1 \ldots x_m] \propto \int \Theta \cdot \Theta^{\#HEADS} (1 - \Theta)^{\#TAILS} d\Theta \)
  
  - Solution (integration by parts):
    
    - \( \Pr[HEADS|x_1 \ldots x_m] = \frac{\#HEADS+1}{m+2} \)
Difficulties in Bayes Reasoning

• Often $\Pr[x|x_1 \ldots x_m]$ has no closed form

• Possible solutions:
  • Can try to find a point estimate $\Theta$
    • MAP (Maximum A-Posteriori)
      • Take $\hat{\Theta} = \arg\max_{\Theta} \Pr[\Theta|x_1 \ldots x_m]$ (The distribution “mode”)
      • Note: This is not the same as MLE!
    • Take $\hat{\Theta} = E[\Theta|x_1 \ldots x_m]$  
    • Take $\hat{\Theta} = median[\Theta|x_1 \ldots x_m]$
MAP

• Advantage:
  • An optimization problem – Often easier to solve

• Disadvantages:
  • Imposes “fake certainty”
    • Unlike mean/median, “Mode” doesn’t mean “typical” (especially in skewed distributions)
  • Sensitive to parametrization
    • If $\Psi = f(\Theta)$, $\hat{\Psi}$ a MAP for $P[\Psi]$, $\hat{\Theta}$ a MAP for $P[\Theta]$, then in general it is not true that $\hat{\Psi} = f(\hat{\Theta})$
Summary

• Generative vs. Discriminative
  • ERM is discriminative, often probabilistic
  • Generative allows us to leverage $\Pr[X]$ (if we trust it)
    • Parametric families simplify model, reduce number of parameters
    • MLE approach finds $\arg\max_{\theta} \Pr[\text{observation}]$
  • Discriminative concentrates on $\Pr[Y|X]$ which is often what we really care about. More appropriate if we don’t trust our knowledge of $\Pr[X]$

• Naïve Bayes yields a linear predictor
  • The “generative version” of Linear Regression

• LDA yields a linear predictor
  • Requires same covariance for all class conditionals
  • Unlike Naïve Bayes, doesn’t require diagonal covariance (=cond. Indep.)

• Some models introduce latent variables

• Bayes reasoning assumes prior “belief” on parameters, predicts by updating “belief” based on observations
  • Often difficult to compute in closed form
  • Point estimates can be used instead
    • MAP/expectation/median