CS340: Machine Learning

Review for midterm

Kevin Murphy

Outline

- Probability basics
- Probability density functions
- Parameter estimation
- Prediction
- Multivariate probability models
- Conditional probability models
- Information theory

Machine learning: a probabilistic approach

- We want to make models of data so we can find patterns and predict the future.
- We will use probability theory to represent our uncertainty about what the right model is; this will induce uncertainty in our predictions.
- We will update our beliefs about the model as we acquire data (this is called "learning"); as we get more data, the uncertainty in our predictions will decrease.

Probability basics

- Chain (product) rule

$$p(A, B) = p(A)p(B|A) \tag{1}$$

- Marginalization (sum) rule

$$p(B) = \sum_a p(A = a, B) \tag{2}$$

- Hence we derive Bayes rule

$$p(A|B) = \frac{p(A, B)}{p(B)} \tag{3}$$

PMF

- A probability mass function (PMF) is defined on a discrete random variable $X \in \{1, \ldots, K\}$ and satisfies

$$1 = \sum_{x=1}^{K} P(X = x) \tag{4}$$

$$0 \leq P(X = x) \leq 1 \tag{5}$$

- This is just a histogram!

PDF

- A probability density function (PDF) is defined on a continuous random variable $X \in \mathbb{R}$ or $X \in \mathbb{R}^+$ or $X \in [0, 1]$ etc and satisfies

$$1 = \int_{\mathcal{X}} p(X = x) dx \tag{6}$$

$$0 \leq p(X = x) \tag{7}$$

- Note that it is possible that $p(X = x) > 1$, since multiplied by dx.
- $P(x \leq X \leq x + dx) \approx p(x)dx$
- Statistics books often write $f_X(x)$ for pdfs and use $P()$ for pmf's.

Moments of a distribution

- Mean (first central moment)

$$\mu = E[X] = \sum_x x p(x) \tag{8}$$

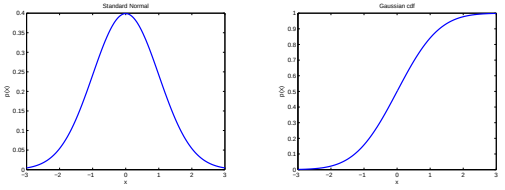
- Variance (second central moment)

$$\sigma^2 = \text{Var}\,[X] = \sum_x (x - \mu)^2 p(x) = E[X^2] - \mu^2 \tag{9}$$

CDFs

The cumulative distribution function is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x) dx \tag{10}$$



Quantiles of a distribution

$$P(a \leq X \leq b) = F_X(b) - F_X(a) \tag{11}$$

eg if $Z \sim \mathcal{N}(0,1)$, $F_Z(z) = \Phi(z)$. Then

$$p(-1.96 \leq Z \leq 1.96) = \Phi(1.96) - \Phi(-1.96) \tag{12}$$

$$= 0.975 - 0.025 = 0.95 \tag{13}$$

Hence $X \sim \mathcal{N}(\mu, \sigma)$,

$$p(-1.96\sigma < X - \mu < 1.96\sigma) = 1 - 2 \times 0.025 = 0.95 \tag{14}$$

Hence the interval $\mu \pm 1.96\sigma$ contains 0.95 mass.

Outline

- Probability basics $\surd$
- Probability density functions
- Parameter estimation
- Prediction
- Multivariate probability models
- Conditional probability models
- Information theory

Bernoulli distribution

- $X \in \{0, 1\}$, $p(X = 1) = \theta$, $p(X = 0) = 1 - \theta$
- $X \sim Be(\theta)$
- $E[X] = \theta$

Multinomial distribution

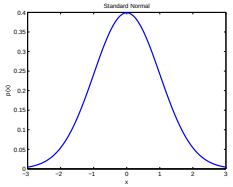
- $X \in \{1, \ldots, K\}$, $p(X = k) = \theta_k$, $X \sim Mu(\theta)$
- $0 \leq \theta_k \leq 1$
- $\sum_{k=1}^{K} \theta_k = 1$
- $E[X_i] = \theta_i$

(Univariate) Gaussian distribution

- $X \in \mathbb{R}$, $X \sim \mathcal{N}(\mu, \sigma)$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$

$$\mathcal{N}(x|\mu, \sigma) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{15}$$

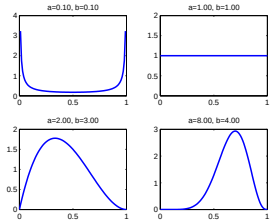
- $E[X] = \mu$



Beta distribution

- $X \in [0, 1]$, $X \sim Be(\alpha_1, \alpha_0)$, $0 \leq \alpha_0, \alpha_1$
- $E[X] = \frac{\alpha_1}{\alpha_1 + \alpha_0}$

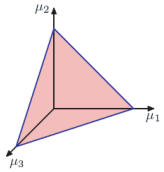
$$Be(X|\alpha_1, \alpha_0) \propto [X^{\alpha_1 - 1}(1 - X)^{\alpha_0 - 1}]$$

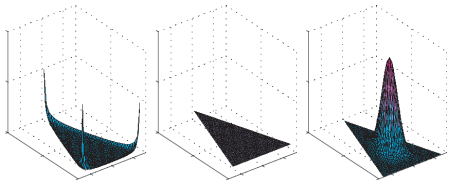


Dirichlet distribution

- $X \in [0, 1]^K$, $X \sim Dir(\alpha_1, \ldots, \alpha_K)$, $0 \leq \alpha_k$
- $E[X_i] = \frac{\alpha_i}{\alpha}$, where $\alpha = \sum_k \alpha_k$.

$$Dir(X|\alpha_1, \ldots, \alpha_K) \propto \prod_k X_k^{\alpha_k - 1}$$





Outline

- Probability basics $\surd$
- Probability density functions $\surd$
- Parameter estimation
- Prediction
- Multivariate probability models
- Conditional probability models
- Information theory

Parameter estimation: MLE

- Likelihood of iid data $D = \{x_n\}_{n=1}^N$

$$L(\theta) = p(D|\theta) = \prod_n p(x_n|\theta) \tag{16}$$

- Log likelihood

$$\ell(\theta) = \log p(D|\theta) = \sum_n \log p(x_n|\theta) \tag{17}$$

- MLE = Maximum likelihood estimate

$$\hat{\theta} = \arg\max_\theta L(\theta) = \arg\max_\theta \ell(\theta) \tag{18}$$

hence

$$\frac{\partial L}{\partial \theta_i} = 0 \tag{19}$$

MLE for Bernoullis

$$p(X_n|\theta) = \theta^{X_n}(1-\theta)^{1-X_n} \tag{20}$$

$$\ell(\theta) = \sum_n X_n \log\theta + \sum_n (1-X_n)\log(1-\theta) \tag{21}$$

$$= N_1 \log\theta + N_0 \log(1-\theta) \tag{22}$$

$$N_1 = \sum_n I(X_n = 1) \tag{23}$$

$$\frac{dL}{d\theta} = \frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0 \tag{24}$$

$$\hat{\theta} = \frac{N_1}{N_1 + N_0} \tag{25}$$

MLE for Multinomials

$$p(X_n|\theta) = \prod_k \theta_k^{I(X_n=k)} \tag{26}$$

$$\ell(\theta) = \sum_n \sum_k I(x_n = k)\log\theta_k \tag{27}$$

$$= \sum_k N_k \log\theta_k \tag{28}$$

$$\tilde{l} = \sum_k N_k \log\theta_k + \lambda\left(1 - \sum_k \theta_k\right) \tag{29}$$

$$\frac{\partial \tilde{l}}{\partial \theta_k} = \frac{N_k}{\theta_k} - \lambda = 0 \tag{30}$$

$$\hat{\theta}_k = \frac{N_k}{N} \tag{31}$$

Parameter estimation: Bayesian

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{32}$$

$$= [\prod_n p(x_n|\theta)]p(\theta) \tag{33}$$

We say $p(\theta)$ is conjugate to $p(D|\theta)$ if $p(\theta|D)$ in same family as $p(\theta)$. This allows recursive (sequential) updating.

MLE is a point estimate. Bayes computes full posterior, natural model of uncertainty (e.g., posterior variance, entropy, credible intervals, etc)

Parameter estimation: MAP

For many models, as $|D| \to \infty$,

$$p(\theta|D) \to \delta(\theta - \hat{\theta}_{MAP}) \to \delta(\theta - \hat{\theta}_{MLE}) \tag{34}$$

where

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(D|\theta)p(\theta) \tag{35}$$

MAP (maximum a posteriori) estimate is a posterior mode. Also called a penalized likelihood estimate since

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \log p(D|\theta) - \lambda c(\theta) \tag{36}$$

where $c(\theta) = -\log p(\theta)$ is a penalty or regularizer, and λ is the strenghth of the regularizer.

Bayes estimate for Bernoulli

$$\begin{aligned}
X_n &\sim Be(\theta) && (37)\\
\theta &\sim Beta(\alpha_1, \alpha_0) && (38)\\
p(\theta|D) &\propto [\prod_n \theta^{X_n}(1-\theta)^{1-X_n}]\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} && (39)\\
&= \theta^{N_1}(1-\theta)^{N_0}\theta^{\alpha_1-1}(1-\theta)^{\alpha_0-1} && (40)\\
&\propto Beta(\alpha_1 + N_1, \alpha_0 + N_0) && (41)
\end{aligned}$$

Bayes estimate for Multinomial

$$\begin{aligned}
X_n &\sim Mu(\theta) && (42)\\
\theta &\sim Dir(\alpha_1, \ldots, \alpha_K) && (43)\\
p(\theta|D) &\propto [\prod_n \prod_k \theta_k^{I(X_n=k)}] \prod_k \theta_k^{\alpha_k-1} && (44)\\
&= \prod_k \theta_k^{N_k}\theta_k^{\alpha_k-1} && (45)\\
&\propto Dir(\alpha_1 + N_1, \ldots, \alpha_K + N_K) && (46)
\end{aligned}$$

Outline

- Probability basics $\surd$
- Probability density functions $\surd$
- Parameter estimation $\surd$
- Prediction
- Multivariate probability models
- Conditional probability models
- Information theory

Predicting the future

Posterior predictive density gotten by Bayesian model averaging

$$p(X|D) = \int p(X|\theta)p(\theta|D)d\theta \tag{47}$$

Plug-in principle: if $p(\theta|D) \approx \delta(\theta - \hat{\theta})$, then

$$p(X|D) \approx p(X|\hat{\theta}) \tag{48}$$

Predictive for Bernoullis

$$\begin{aligned} p(\theta|D) &= Beta(\alpha_1 + N_1, \alpha_0 + N_0) && (49)\\ &\stackrel{\text{def}}{=} Beta(\alpha_1', \alpha_0') && (50)\\ p(X=1|D) &= \int p(X=1|\theta)p(\theta|D)d\theta && (51)\\ &= \int \theta p(\theta|D)d\theta && (52)\\ &= E[\theta|D] && (53)\\ &= \frac{\alpha_1'}{\alpha_1' + \alpha_0'} && (54) \end{aligned}$$

Cross validation

We can use CV to find the model with best predictive performance.

Outline

- Probability basics $\surd$
- Probability density functions $\surd$
- Parameter estimation $\surd$
- Prediction $\surd$
- Multivariate probability models
- Conditional probability models
- Information theory

Multivariate probability models

- Each data case X_n is now a vector of variables X_{ni}, $i = 1 : p$, $n = 1 : N$
- e.g., $X_{ni} \in \{1, \ldots, K\}$ is the i'th word in n'th sentence
- e.g., $X_{ni} \in \mathbb{R}$ is the i'th attribute in the n'th feature vector
- We need to define $p(X_n|\theta)$ for feature vectors of potentially variable size.

Independent features (bag of words model)

$$p(X_n|\theta) = \prod_i p(X_{ni}|\theta_i) \tag{55}$$

e.g., $X_{ni} \in \{0, 1\}$, product of Bernoullis

$$p(X_n|\theta) = \prod_i \theta_i^{X_{ni}} (1 - \theta_i)^{X_{ni}} \tag{56}$$

e.g., $X_{ni} \in \{1, \ldots, K\}$, product of multinomials

$$p(X_n|\theta) = \prod_i \prod_k \theta_{ik}^{I(X_{ni}=k)} \tag{57}$$

MLE for bag of words model

$$p(D|\theta) = \prod_n \prod_i p(X_{ni}|\theta_i) \tag{58}$$

We can estimate each θ_i separately, eg. product of multinomials

$$p(D|\theta) = \prod_n \prod_i \prod_k \theta_{ik}^{I(X_{ni}=k)} \tag{59}$$

$$= \prod_i \prod_k \theta_{ik}^{N_{ik}} \tag{60}$$

$$N_{ik} \stackrel{\text{def}}{=} \sum_n I(X_{ni} = k) \tag{61}$$

$$\ell(\theta_i) = \sum_k N_{ik} \log \theta_{ik} \tag{62}$$

$$\theta_{ik} = \frac{N_{ik}}{N_i} \tag{63}$$

Bayes estimate for bag of words model

$$p(\theta|D) \propto \prod_i p(\theta_i)[\prod_n \prod_i p(X_{ni}|\theta_i)] \tag{64}$$

We can update each θ_i separately, eg. product of multinomials

$$p(\theta|D) = \prod_i Dir(\theta_i|\alpha_{i1}, \ldots, \alpha_{iK})[\prod_k \theta_{ik}^{N_{ik}}] \tag{65}$$

$$\propto \prod_i Dir(\theta_i|\alpha_{i1} + N_{i1}, \ldots, \alpha_{iK} + N_{iK}) \tag{66}$$

Factored prior $\times$ factored likelihood = factored posterior

(Discrete time) Markov model

$$p(X_n|\theta) = p(X_{n1}|\pi)\prod_{t=2}^{p} p(X_{nt}|X_{n,t-1}, A) \tag{67}$$

For a discrete state space, $X_{nt} \in \{1, \ldots, K\}$,

$$\pi_i = p(X_1 = i) \tag{68}$$

$$A_{ij} = p(X_t = j|X_{t-1} = i) \tag{69}$$

$$p(X_n|\theta) = \prod_i \pi_i^{I(X_{n1}=i)} \prod_t \prod_i \prod_j A_{ij}^{I(X_{nt}=j, X_{n,t-1}=i)} \tag{70}$$

$$= \prod_i \pi_i^{N_i^1} \prod_i \prod_j A_{ij}^{N_{ij}} \tag{71}$$

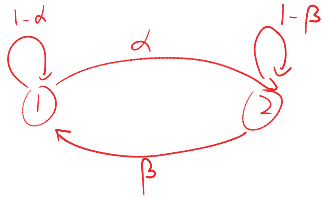
$$N_i^1 \stackrel{\text{def}}{=} I(X_{n1} = i) \tag{72}$$

$$N_{ij} \stackrel{\text{def}}{=} \sum_t I(X_{n,t} = j, X_{n,t-1} = i) \tag{73}$$

Stationary distribution for a Markov chain

π_i is fraction of time we spend in state i, given by principle eigenvector

$$A\pi = \pi \tag{80}$$



Balance flow across cut set

$$\pi_1\alpha = \pi_2\beta \tag{81}$$

Since $\pi_1 + \pi_2 = 1$, we have

$$\pi_1 = \frac{\beta}{\alpha+\beta}, \quad \pi_2 = \frac{\alpha}{\alpha+\beta}, \tag{82}$$

MLE for Markov model

$$p(D|\theta) = \prod_n \prod_i \pi_i^{I(X_{n1}=i)} \prod_t \prod_i \prod_j A_{ij}^{I(X_{nt}=j, X_{n,t-1}=i)} \tag{74}$$

$$= \prod_i \pi_i^{N_i^1} \prod_i \prod_j A_{ij}^{N_{ij}} \tag{75}$$

$$N_i^1 \stackrel{\text{def}}{=} \sum_n I(X_{n1} = i) \tag{76}$$

$$N_{ij} \stackrel{\text{def}}{=} \sum_n \sum_t I(X_{n,t} = j, X_{n,t-1} = i) \tag{77}$$

$$\hat{A}_{ij} = \frac{N_{ij}}{\sum_{j'} N_{ij'}} \tag{78}$$

$$\hat{\pi}_i = \frac{N_i^1}{N} \tag{79}$$

Outline

- Probability basics $\surd$
- Probability density functions $\surd$
- Parameter estimation $\surd$
- Prediction $\surd$
- Multivariate probability models $\surd$
- Conditional probability models
- Information theory

Conditional density estimation

- Goal: learn $p(y|x)$ where x is input and y is output.
- Generative model

$$p(y|x) \propto p(x|y)p(y) \tag{83}$$

- Discriminative model e.g., logistic regression

$$p(y|x) = \sigma(w^T x) \tag{84}$$

Naive Bayes = conditional bag of words model

Generative model

$$p(y|x) \propto p(x|y)p(y) \tag{85}$$

in which features of x are conditionally independent given y:

$$p(y|x) \propto p(y)[\prod_{i=1}^{p} p(x_i|y)] \tag{86}$$

e.g., for multinomials

$$p(y|x) \propto \prod_c \pi_c^{I(y=c)} [\prod_i \prod_k \theta_{cik}^{I(x_i=k, y=c)}] \tag{87}$$

MLE for Naive Bayes

We estimate params for each feature separately, partitioning the data based on label y.

$$p(D|\theta, \pi) = [\prod_c \pi_c^{N_c}] \prod_i [\prod_k \prod_c \theta_{cik}^{N_{cik}}] \tag{88}$$

$$N_{cik} = \sum_n I(x_{ni} = k, y_n = c) \tag{89}$$

$$\hat{\theta}_{cik} = \frac{N_{cik}}{\sum_{k'} N_{cik'}} \tag{90}$$

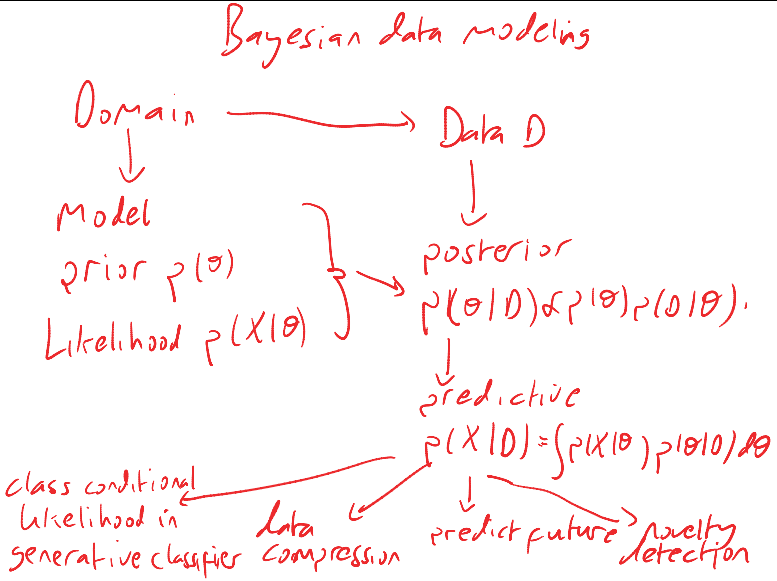
$$\hat{\pi}_c = \frac{N_c}{N} \tag{91}$$

Conditional Markov model

$$p(x|y) = p(x_1|y) \prod_t p(x_t|x_{t-1}, y) \tag{92}$$

Fit a separate Markov model for every class of y. See homework 5.

Summary



Summary

- Probability basics: Bayes rule, pdfs, pmfs
- Probability density functions: Bernoulli, Multinomial; Gaussian, Beta, Dirichlet
- Parameter estimation: MLE, Bayesian
- Prediction: Plug-in, Bayesian, cross validation
- Multivariate probability models: Bag of words, Markov models
- Conditional probability models: Conditional BOW (Naive Bayes), Conditional Markov
- Information theory

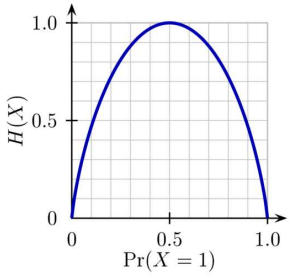
Entropy

- Measure of uncertainty.
- Definition for PMF p_k, $k = 1 : K$

$$H(p) = -\sum_k p_k \log_2 p_k \tag{93}$$

- eg binary entropy function, $p_1 = \theta$, $p_0 = 1 - \theta$

$$H(\theta) = -[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)] \tag{94}$$
$$= -[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)] \tag{95}$$



Joint entropy

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y) \tag{96}$$

If $X \perp Y$, then $H(X,Y) = H(X) + H(Y)$.
Pf:

$$H(X,Y) = -\sum_{x,y} p(x)p(y)\log p(x)p(y) \tag{97}$$
$$= -\sum_{x,y} p(x)p(y)\log p(x) - \sum_{x,y} p(x)p(y)\log p(y) \tag{98}$$

In general

$$H(X,Y) \leq H(X) + H(Y) \tag{99}$$

Pf:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \geq 0 \tag{100}$$

Conditional entropy

$$H(Y|X) \stackrel{\text{def}}{=} \sum_x p(x) H(Y|X=x) \quad (101)$$

$$= H(X,Y) - H(X) \quad (102)$$

In hw 5, you showed that $H(Y|X) \leq H(Y)$.
Pf

$$H(Y|X) = H(X,Y) - H(X) \leq H(X) + H(Y) - H(Y) \quad (103)$$

But $\exists y.H(X|y) > H(X)$.

Mutual information

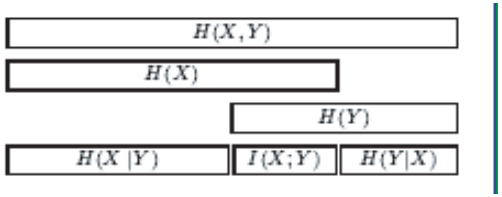
$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \quad (104)$$

In hw 5, you showed that

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \quad (105)$$

Subsituting $H(Y|X) = H(X,Y) - H(X)$ yields

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \quad (106)$$



Mutual information ≥ 0

$I(X,Y) = 0$ if $X \perp Y$. Pf.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x)p(y) \log \frac{p(x)p(y)}{p(x)\,p(y)} \quad (107)$$

$$= \sum_y \sum_x p(x)p(y) \log 1 = 0 \quad (108)$$

$I(X,Y) \geq 0$. Pf.

$$I(X,Y) = KL(p(x,y)||p(x)p(y)) \geq 0 \quad (109)$$

Relative entropy

$$\mathcal{D}(p||q) \stackrel{\text{def}}{=} \sum_k p_k \log \frac{p_k}{q_k} \quad (110)$$

$$= \sum_k p_k \log p_k - \sum_k p_k \log q_k \quad (111)$$

$$= -\sum_k p_k \log q_k - H(p) \quad (112)$$

Hence measures extra number of bits needed to encode data if we use q instead of p.
$D(p||q) = 0$ if $p = q$. Pf: trivial.
$D(p||q) \geq 0$. Pf: use Jensen's inequality.

MDL principle

Minimum description length principle.
To encode an event $X = x$ which occurs with $p(x)$ takes $-\log p(x)$ bits. Total cost of data D and model H is

$$L(D, H) = -\log p(D|H) - \log p(H) \tag{113}$$

MDL principle says: pick

$$H^* = \arg\min_H L(D, H) \tag{114}$$

This is equivalent to the MAP hypothesis

$$H^* = \arg\max_H \log p(D|H) + \log p(H) \tag{115}$$