Reading: Ng's course: cs229-prob.pdf

Probability Review

- Based on experiment: outcome space Ω containing all possible atomic outcomes
- Each outcome (atom) has probability density or mass (discrete vs. continuous spaces)
- Event is a subset of Ω
- P(event) is sum (or integral) over event's atoms
- Random variable V maps Ω to (usually) R
- V=value is an event, <u>P(V) is a distribution</u>

Example

- Roll a fair 6-sided die and then flip that many fair coins.
- What is Ω?

Example

- Roll a fair 6-sided die and then flip that many fair coins.
- What is Ω?
- Ω={(1,H), (1,T), (2, HH), (2, HT), …, (6,TTTTTT)}
- Number of heads is a random variable
- What is expected number of heads?

Expectation of *V* is $\sum_{\text{atoms } a} P(a)V(a)$

- Events A and B independent iff:

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

- Probability of A given B, (cond. probability)

$$P(A \mid B) = P(A \text{ and } B) / P(B)$$

- So,

$$P(A \text{ and } B) = P(A \mid B) \cdot P(B)$$

$$P(B \text{ and } A) = P(B \mid A) \cdot P(A)$$

Product rule

- Bayes Rule:

$$P(A \mid B) = P(B \mid A)\, P(A) / P(B)$$

Example

What is expected number of heads?

def. expectation: $E(V) = \sum_{\text{atoms } a} P(a)V(a)$

- Expectations add: $E(V_1 + V_2) = E(V_1) + E(V_2)$
- Rule of conditioning: (sum rule)

if events $e_1, e_2, \ldots, e_k$ **partition** Ω then:

P(event) = $\sum$ P(e_i) P(event | e_i) = $\sum$ P(e_i and event)

E(randVar) = $\sum$ P(e_i) E(randVar | e_i)

Expected number of heads

- E(# heads) $= \sum_{r=1}^{6} P(\text{roll} = r) E(\#\,\text{heads} \mid \text{roll} = r)$

$$= \frac{1}{6}\left(\frac{1+2+3+4+5+6}{2}\right)$$

$$= \frac{21}{12} = 1.75$$

- Joint Distributions factor:

  If $\Omega = (S \times T \times U)$ then $P(S=s, T=t, U=u)$ is

  $P(S=s)\, P(T=t \mid S=s)\, P(U=u \mid S=s, T=t)$

  (can draw one at a time with conditioning)

- Conditional distributions are distributions:

  $P(A \mid B) = P(A \text{ and } B) / P(B)$, so also:

  $P(A \mid B, C) = P(A \text{ and } B \mid C) / P(B \mid C)$

Bayes Rule for Learning

RVs

- Assume joint distribution $P(\boldsymbol{X}=\boldsymbol{x}, Y=y)$
- Want $P(Y=y \mid \boldsymbol{X}=\boldsymbol{x})$ for each label y on a new instance $\boldsymbol{x}$ (here $(\boldsymbol{x},y)$ is an atom)
- $P(y \mid \boldsymbol{x}) = P(\boldsymbol{x} \mid y) \cdot P(y) / P(\boldsymbol{x})$ Bayes rule
- $P(y \mid \boldsymbol{x})$ proportional to $P(\boldsymbol{x} \mid y) \cdot P(y)$
- From data, learn $P(\boldsymbol{x} \mid y)$ and $P(y)$
- Predict label y with largest product

How to learn probabilities

- Street hustler takes bets on coin flips
- You see HTH, what is probability that next flip is H? What is θ=P(H) for coin?

(don't be shy)

How to learn probabilities

- Street hustler takes bets on coin flips
- You see HTH, what is probability that next flip is H? What is h=P(H) for coin?

- What is the experiment?

Frequentist

- Street hustler takes bets on coin
- You see event HTH, what is probability that next flip is H?
- Frequentist: 2/3 maximizes likelihood, the θ=P(H) value maximizing $\theta^2(1-\theta)$
  solve "derivative = 0" for θ.
- Likelihood function $L(\theta) = P(HTH \mid \theta)$
  vs. the probability $P(HTH \mid \theta=2/3)$

Bayesian Parameter Estimation

- Have prior distribution $P(\theta)$ on $\theta = P(H)$; two phase experiment, pick θ then flip 3 times
- Posterior on θ is distribution $P(\theta \mid HTH)$ or

$$P(HTH \mid \theta) \cdot P(\theta) \,/\, P(HTH)$$

- In this case,

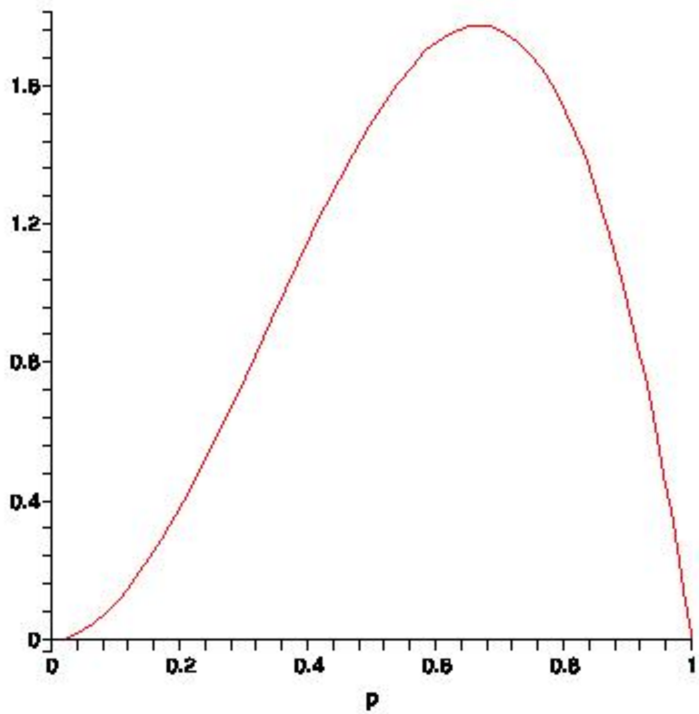
$$\theta^2(1-\theta)P(\theta) \,/\, \text{normalization}$$

Bayesian examples

- Prior: $P(\theta=0) = P(\theta=1/2) = P(\theta=1) = 1/3$:
  $\theta^2(1-\theta)\ P(\theta)$ is 0, 1/24, and 0 for these three cases, posterior $P(\theta=1/2 \mid HTH) = 1$

- Prior density: $P(\theta) = 1$ for $0 \le \theta \le 1$:
  $\theta^2(1-\theta)\ P(\theta)$ is $\theta^2(1-\theta)$ for $0 \le \theta \le 1$
  posterior $P(\theta \mid HTH)$ is $12\ \theta^2(1-\theta)$

Posterior plot



- Max at 2/3
- Average is 3/5
- 3/5 = (2+1)/(3+2)
- Not a coincidence!

  Laplace's rule of succession - add one fictitious observation of each class

Bayes' Estimation

- Treat parameter θ as a random var with the prior distribution $P(\theta)$, use fixed sample $\mathcal{X}$ (RV S)
- Maximum Likelihood (ML):
  - $\theta_{ML} = \text{argmax}_{\theta'} P(S=\mathcal{X} \mid \theta = \theta')$
- Maximum a Posteriori (MAP):
  - $\theta_{MAP} = \text{argmax}_{\theta'} P(\theta=\theta' \mid S=\mathcal{X})$
    $= \text{argmax}_{\theta'} P(S=\mathcal{X} \mid \theta=\theta') \, P(\theta=\theta') \, / \, P(S=\mathcal{X})$
- Predictive distribution (full Bayes):

  $P(Y=y|S=\mathcal{X}) = \int P(Y=y|\theta=\theta') \, P(\theta=\theta'|S=\mathcal{X}) \, d\theta'$

Mean a'Post.: $\theta_{mean} = E[\theta|S=\mathcal{X}] = \int \theta' \, P(\theta=\theta'|S=\mathcal{X}) \, d\theta'$

Use for learning

RVs

- Draw enough data so that $P(Y=y \mid \boldsymbol{X}=\boldsymbol{x})$ estimated for every possible $(\boldsymbol{x},y)$ pair
- This takes lots of data – curse of dimensionality …rote learning
- Another approach: a class of models
- Think of each model as a way of generating the training set $\mathcal{X}$ of $(\boldsymbol{x},y)$ pairs

Compound Experiment

- Prior $P(M=m)$ on model space
- Models gives $P(X=\mathcal{X} \mid M=m)$

(here data $\mathcal{X}$ is both y's and $\boldsymbol{x}$'s)

Joint experiment (if data i.i.d. given m)

- $P(\{(\boldsymbol{x}_i, y_i)\}, m) = P(m) \prod_i (P(\boldsymbol{x}_i|m) P(y_i \mid \boldsymbol{x}_i, m))$

This is a generative model – has $P((x,y) \mid m)$

- **Prior** $P(m)$ over models
- Model gives $P(\mathcal{X} \mid m)$
- **Posterior** $P(m \mid \mathcal{X}) = P(\mathcal{X} \mid m)\, P(m) \,/\, P(\mathcal{X})$
- *Max. likelihood*: m having max $P(\mathcal{X} \mid m)$
- *Max. a'posteriori*: m having max $P(m \mid \mathcal{X})$
- *Predictive distribution (full Bayes)*: predict with average of m's weighted by posteriors $P(m \mid \mathcal{X})$

Discriminative and Generative models

- **Generative model**: $P((\boldsymbol{x}, y) \mid m)$
  - Tells how to generate examples (both instances and labels)
- **Discriminative model**: $P(y \mid h, \boldsymbol{x})$
  - Tells how to create labels from instances, (like linear regression)
- **Discriminate function**: predict $f(\boldsymbol{x})$, often $f(\boldsymbol{x}) = \text{argmax}_t \ f_t(\boldsymbol{x})$.

More on Generative approach

- Generative approach models $P(\boldsymbol{x}, y \mid m)$
- Learn $P(\boldsymbol{x} \mid y, m)$ and use Bayes' rule

  $P(y \mid \boldsymbol{x}, m) = P(\boldsymbol{x} \mid y, m)\, P(y \mid m) \,/\, P(\boldsymbol{x} \mid m)$
- Need model for $P(\boldsymbol{x} \mid y, m)$
- One common assumption:

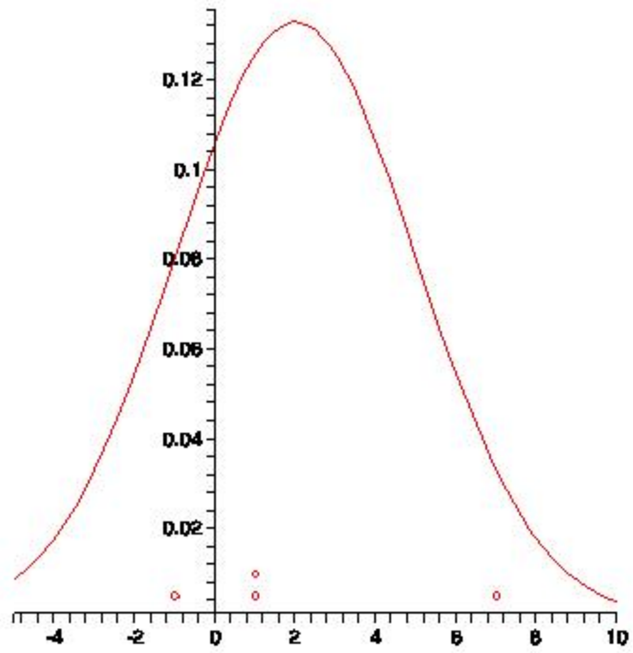
  $P(\boldsymbol{x} \mid y, m)$ Gaussian

  $P(y \mid m)$ Bernoulli (biased coin flip)
- How to learn (fit) Gaussian from data?

1 dimensional Gausians

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- Maximum likelihood estimate has sample mean μ and sample variance $\sigma^2 = (1/n) \sum_i (x_i - \mu)^2$
$= E[(x - \mu)^2]$
$= E[x^2] - \mu^2$
- What gaussian best fits -1, 1, 1, 7?



0.12
0.1
0.08
0.06
0.04
0.02
-4
-2
0
2
4
6
8
10

Multivariate Gaussians

$$P(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2} \mid \Sigma \mid^{1/2}} \exp(-\frac{1}{2}[\boldsymbol{x} - \mu]^T \Sigma^{-1} [\boldsymbol{x} - \mu])$$

- Mean vector μ, covariance matrix Σ, entries of covariance matrix are variances (or co-variances), $\sigma^2_{i,j}$=E[(x_i- μ_i) (x_j- μ_j)] (subscripts are indices into vectors)

Estimating Gaussians
(maximum likelihood)

- Estimate $\boldsymbol{\mu} = (\sum_i \boldsymbol{x_i}) / n$
- Estimate $\sigma^2_{i,j} = (\sum_k (x_{k,i} - \mu_i)(x_{k,j} - \mu_j)) / n$
- (above covariance estimate is <u>biased</u>: can use "*n*-1")
- If domain *d*-dimensional:
  - *d* parameters for $\boldsymbol{\mu}$
  - $d(d+1)/2$ parameters for $\sigma^2_{i,j}$'s
  - For each class!
  - Many parameters "requires" lots of data

Common tricks

- Share same Σ for all classes
  - Gaussian Discriminant Analysis in NG
- Assume diagonal Σ's for each class
- Assume shared $\Sigma = cI$ (spherical)
  - This leads to the simple mean-based linear classifier if data balanced

Gaussian Conditionals and Marginals

- If $p(x,y)$ is Gaussian, then:
  - Conditional $p(x \mid Y=y)$ is Gaussian,
  - Marginal $p(x) = \int p(x,y)dy$ is Gaussian

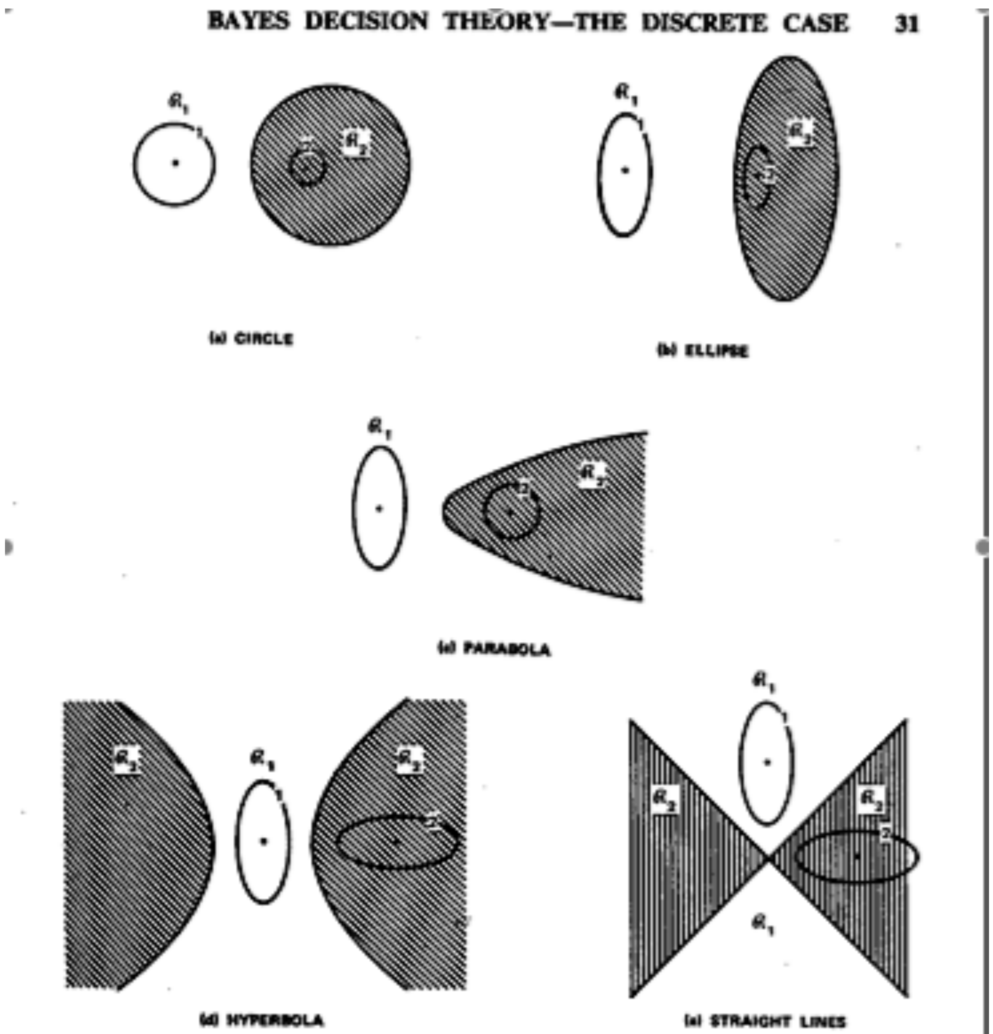


FIGURE 2.10. Forms for decision boundaries for the general bivariate normal case.

General decision boundaries for Gaussian generative model

Duda and Hart' 73

Also: same covar. Matrix implies linear boundary

Main Points/terms:

- Features, instances, labels, examples
- Batch learning, inductive bias
- Classification, regression, loss function
- Training set, training error, test error
- Noise, over-fitting
- Probability: events and RV's, independence, sum rule, product rule, Bayes rule

Main points/terms (cont.)

- Bayesian parameter estimation: priors and posteriors, max.likelihood, max a'posteriori, mean a'posteriori, full Bayesian prediction (predictive distribution)
- Generative models and discriminative models
- Class-conditional gaussians and estimating Gaussians