ISTA 421 + INFO 521
Introduction to Machine Learning

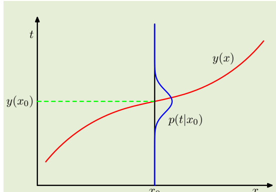
Lecture 11:
The Bayesian Way

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

27 September 2017

1

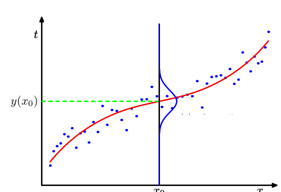
The Maximum Likelihood Way



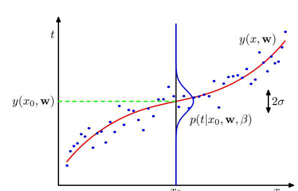
The generating process...

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$



... generates data ...



... that we fit a model to

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

prediction

estimated parameters

Maximum Likelihood Estimates of Params

$\hat{\mathbf{w}}$: $\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$

$\hat{\sigma}^2$: $\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}})$

The MLE is unique

$\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$

Analysis of Uncertainty in Param Estimates via Expectation

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\, \mathbf{w}, \sigma^2)}\left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$

$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}}\right)^{-1}$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$

The Fisher Information $\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$.

Predictions: $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$ $\quad \sigma^2_{\text{new}} = \sigma^2 \mathbf{x}^\mathsf{T}_{\text{new}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$ $\quad \sigma^2_{\text{new}} = \mathbf{x}^\mathsf{T}_{\text{new}}\text{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\text{new}}$

2

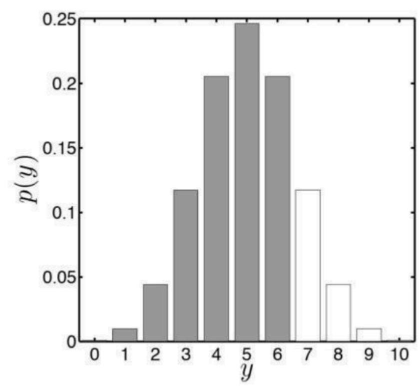
The Coin Game

Place $1 bet
Flip coin 10 times
6 or fewer heads, you win your $1 + $1
More than 6, you loose your $1

Binomial Distribution

$$P(Y = y) = \binom{N}{y} r^y (1 - r)^{N-y}$$

Assume it's a fair coin,
what is your probability of winning?



$$\begin{aligned} P(Y \le 6) = 1 - P(Y > 6) &= 1 - [P(Y = 7) + P(Y = 8) + P(Y = 9) \\ &\quad + P(Y = 10)] \\ &= 1 - [0.1172 + 0.0439 + 0.0098 + 0.0010] \\ &= 0.8281. \end{aligned}$$

3

The Coin Game

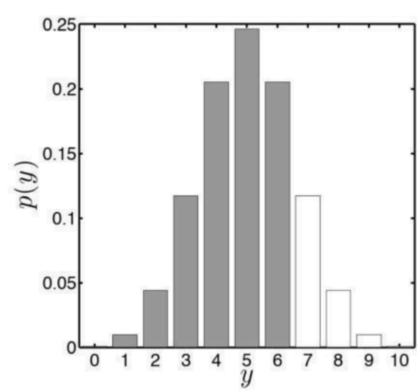
Place $1 bet
Flip coin 10 times
6 or fewer heads, you win your $1 + $1
More than 6, you loose your $1

Binomial Distribution

$$P(Y = y) = \binom{N}{y} r^y (1 - r)^{N-y}$$

What is the expected return (value) from playing the game?

$$\mathbf{E}_{P(x)} \{f(X)\} = \sum_x f(x) P(x)$$



Let X be a random variable, 1=win and 0=lose: $P(X=1) = P(Y \le 6)$
If X=1, get return of $2, so $f(X=1) = 2$, $f(X=0) = 0$.

$$f(1)P(X = 1) + f(0)P(X = 0) = 2 \times P(Y \le 6) + 0 \times P(Y > 6) = 1.6562$$

Given that it costs $1 to play, then on average, we expect to earn 1.6562 – 1 ≈ 66 cents

4

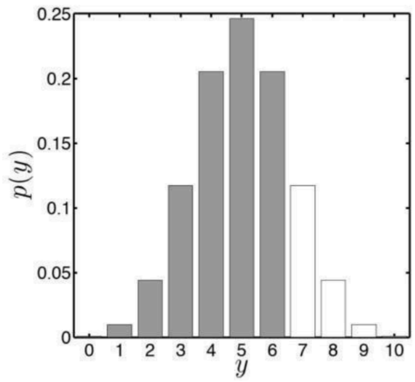
The Coin Game

Place $1 bet
Flip coin 10 times
6 or fewer heads, you win your $1 + $1
More than 6, you loose your $1

Binomial Distribution

$$P(Y=y) = \binom{N}{y} r^y (1-r)^{N-y}$$

Assumptions:
(1) Number of heads is binomial, probability of heads is *r*
(2) The coin is fair: *r* = 0.5



5

Estimate *r* based on evidence

The Maximum Likelihood Way

Observe: H, T, H, H, H, H, H, H, H, H

$$P(Y=y|r,N) = \binom{N}{y} r^y (1-r)^{N-y}$$

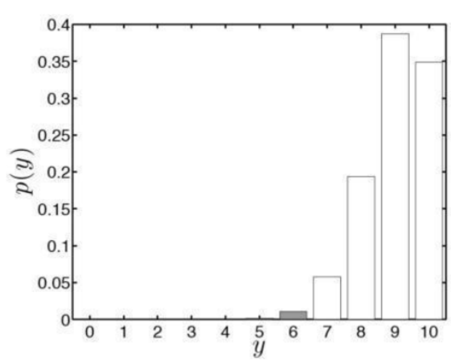
$$L = \log P(Y=y|r,N) = \log \binom{N}{y} + y \log r + (N-y)\log(1-r)$$

$$\begin{aligned} \frac{\partial L}{\partial r} &= \frac{y}{r} - \frac{N-y}{1-r} = 0 \\ y(1-r) &= r(N-y) \\ y &= rN \\ r &= \frac{y}{N}. \end{aligned}$$

r = 0.9, $P(Y \leq 6)$ = 0.0128

$$2 \times P(Y \leq 6) + 0 \times P(Y > 6) = 0.0256$$

Expected value: 0.0256 – 1 = -0.9755



Estimate *r* based on evidence

The Bayesian Way

Observe: H, T, H, H, H, H, H, H, H, H

Think about the specific estimate of *r* as drawn from a random variable *R* – there is inherent uncertainty in our estimate of *r*.

Let random variable Y_N be the number of heads obtained in *N* tosses.

The distribution of *r* conditioned on value of Y_N:

$$p(r|y_N)$$

The expected probability of winning: *the expectation* of $P(Y_{new} \leq 6|r)$ with respect to $p(r|y_N)$

$$P(Y_{\text{new}} \leq 6|y_N) = \int P(Y_{\text{new}} \leq 6|r)p(r|y_N)\text{d}r$$

Random variable representing: The number of heads in a future set of 10 tosses

Estimate *r* based on evidence

The Bayesian Way

Observe: H, T, H, H, H, H, H, H, H, H, H

$$P(Y_{\text{new}} \leq 6|y_N) = \int P(Y_{\text{new}} \leq 6|r)p(r|y_N)\text{d}r$$

We want: $p(r|y_N)$

We have: $p(y_N|r)$

The probability distribution function over the number of heads in *N* independent tosses, where the probability of a head in a single toss is *r*.
This can be represented as the Binomial distribution!

$$P(Y = y) = \binom{N}{y} r^y (1-r)^{N-y}$$

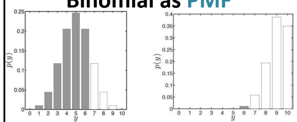
Use Bayes' rule to compute $p(r|y_N)$:

$$p(r|y_N) = \frac{P(y_N|r)p(r)}{P(y_N)}$$

8

Using Bayes' Rule

Binomial as PMF



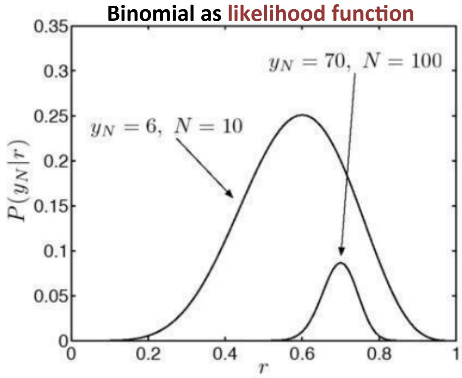
posterior, likelihood, prior, marginal likelihood

$$p(r|y_N) = \frac{P(y_N|r)p(r)}{P(y_N)}$$

(1) **The Likelihood**: $p(y_N|r)$

"How likely is it we would observe our data (y_N) for a particular value of *r* (our model)"

Binomial as likelihood function



$$P(Y = y) = \binom{N}{y} r^y (1-r)^{N-y}$$

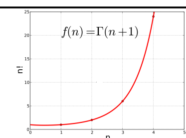
Now we're using the Binomial dist. ***as a function of r***

Remember: Likelihood fn is not itself a probability function!

Both examples tell us different amounts about *r*.

9

Using Bayes' Rule



posterior, likelihood, prior, marginal likelihood

$$p(r|y_N) = \frac{P(y_N|r)p(r)}{P(y_N)}$$

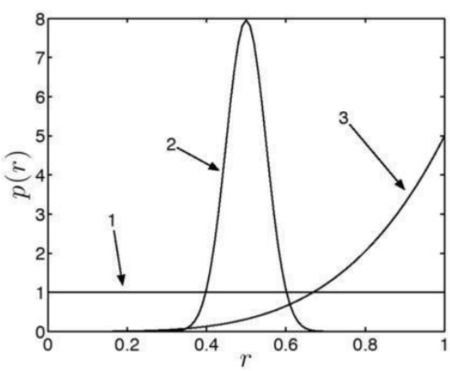
$$\Gamma(n) = (n-1)!$$

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_{r=0}^{r=1} r^{\alpha-1}(1-r)^{\beta-1}\,dr$$

$$\int_{r=0}^{r=1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\,dr = 1$$

(2) **The Prior**: $p(r)$

"Allows us to express any belief we have in the value of *r* ***before*** we see any data."



1) We don't know anything about the coins or the stall owner
$\alpha = 1,\ \beta = 1$

2) We think the coin (and the stall owner) is fair $\alpha = 50,\ \beta = 50$

3) We think the coin (and the stall owner) is biased to flip heads more often
$\alpha = 5,\ \beta = 1$

Beta Distribution

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

Using Bayes' Rule

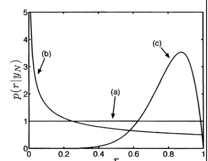
$$\underset{\text{posterior}}{p(r|y_N)} = \frac{\overset{\text{likelihood}}{P(y_N|r)}\overset{\text{prior}}{p(r)}}{\underset{\text{marginal likelihood}}{P(y_N)}}$$

(3) **The Marginal Likelihood**: $P(y_N)$ (aka: the "evidence" or "model evidence")

"Acts as a normalizing constant to ensure $p(r|y_N)$ is a properly defined density."

$$P(y_N) = \int_{r=0}^{r=1} P(y_N|r)p(r)\ dr$$

Known as the **marginal likelihood** because it is the likelihood of the data, y_N, averaged over all parameter values (over all r).



(4) **The Posterior distribution**: $p(r|\ y_N)$

"The result of updating our prior belief $p(r)$ in light of new evidence y_N."

We can use the posterior density to compute expectations

$$\mathbf{E}_{p(r|y_N)}\{P(Y_{10} \leq 6)\} = \int_{r=0}^{r=1} P(Y_{10} \leq 6|r)p(r|y_N)\ dr$$

… the expected value of the probability that we will win!

11

Computing Posteriors

- **Conjugate Prior**: A likelihood-prior pair that results in a posterior which is the same form as the prior

Prior	Likelihood	
Gaussian	Gaussian	μ
Beta	Binomial	
Gamma	Gaussian	$\frac{1}{\sigma^2}$ (precision)
Dirichlet	Multinomial	

$$p(r|y_N) = \frac{\overset{\text{likelihood}}{P(y_N|r)}\overset{\text{prior}}{p(r)}}{P(y_N)}$$

12

Binomial & Beta are Conjugate !

$$P(Y=y) = \binom{N}{y} r^y (1-r)^{N-y}$$

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}$$

$$p(r|y_N) \propto P(y_N|r)p(r)$$

$$p(r|y_N) \propto \left[\binom{N}{y_N} r^{y_N}(1-r)^{N-y_N}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\right]$$

13