10-315
Introduction to ML

Probabilistic Models: MAP

Instructor: Pat Virtue

Announcements

Probability

- Come get help this week

Project

- First step: groups of 3
- See piazza for details

Polls

- Still working on WiFi
  - Please avoid non-315 bandwith
- Mid-semester adjustment
  - 6 free polls: (denominator is currently 22 rather than 28)
- Peer Instruction

Plan

Today

- MLE → MAP

Wed and next Mon

- Neural net applications (image, language)

Next Wed

- Back to probability
- Discriminative → Generative
  - Naïve Bayes
  - Discriminant Analysis
- Combining MAP and Generative

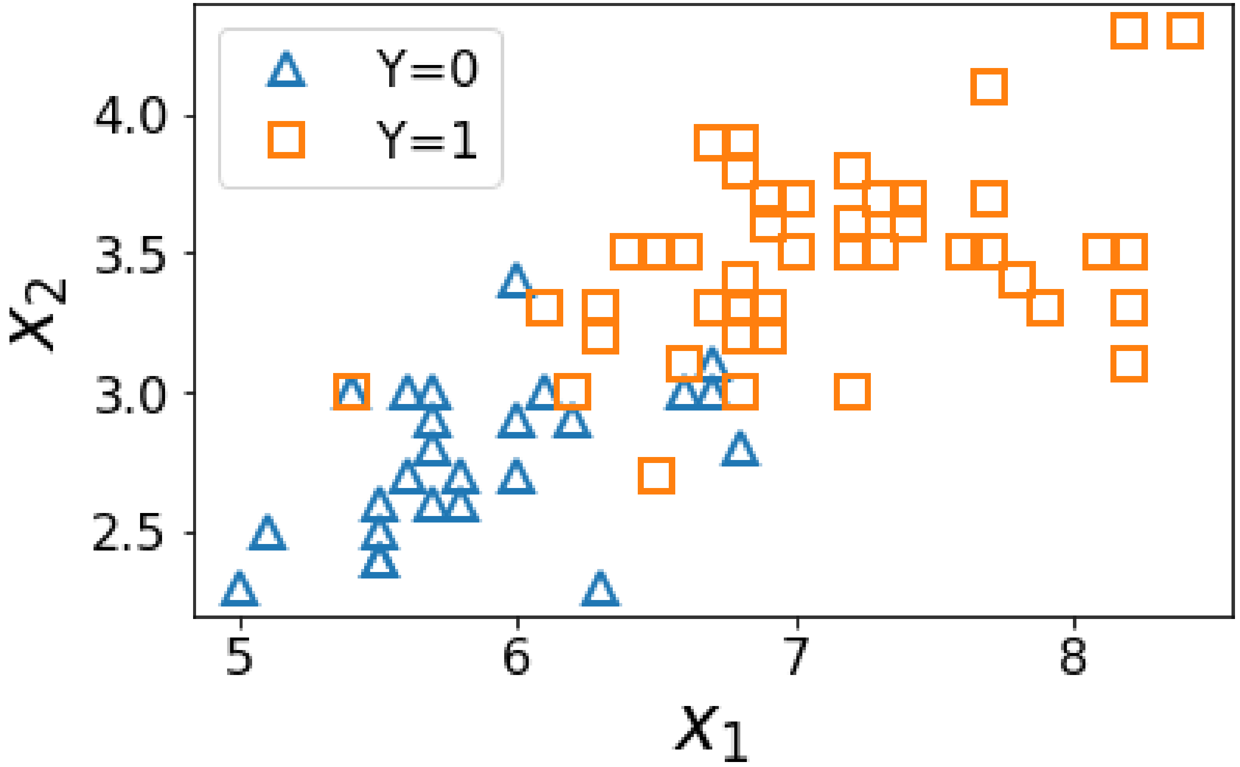
Where are we?

Empirical Risk Minimization vs MLE/MAP

We seem to be redoing a lot of work? ...well, we are. But there is a reason

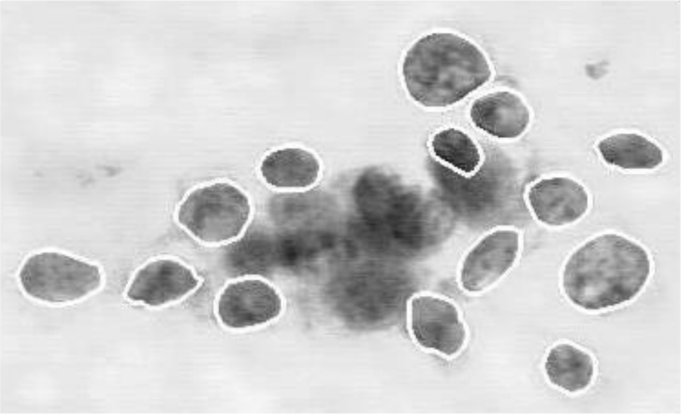
Why Probabilistic Models?

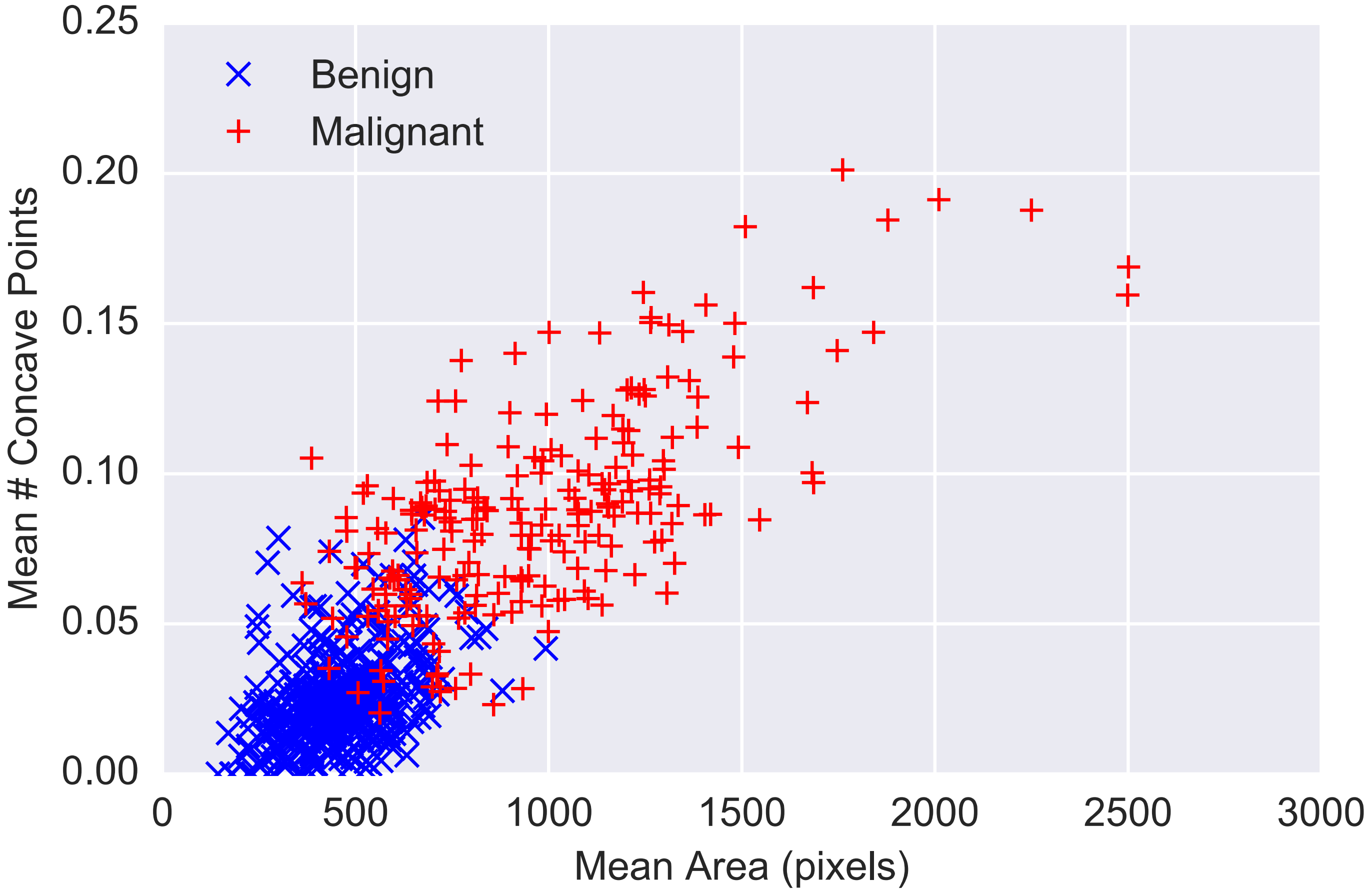
Iris Data



Why Probabilistic Models?

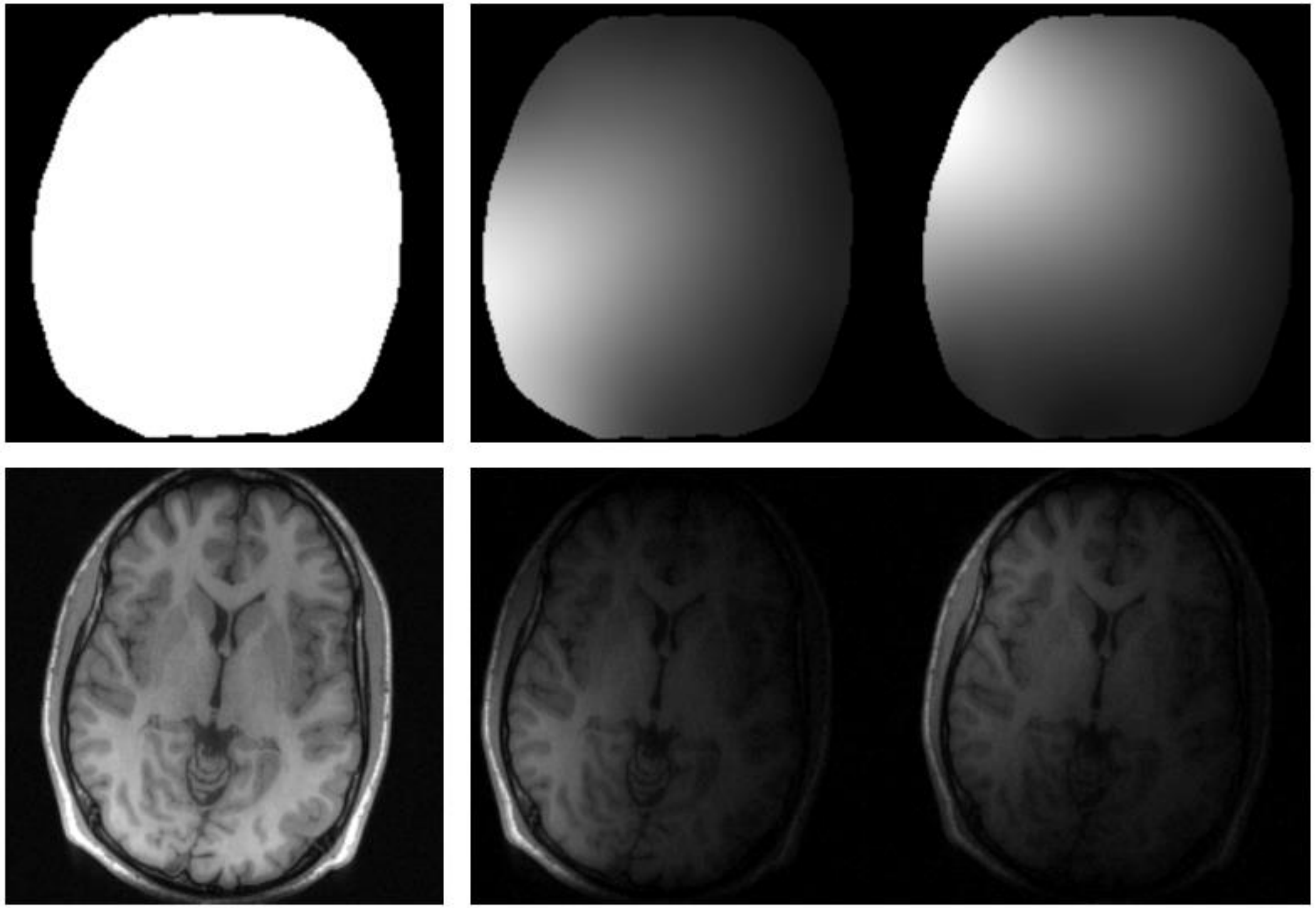
Breast Cancer Diagnosis





Why Probabilistic Models?

MRI Image Reconstruction



Categorical Gaussian Generative Model

PREVIEW

Estimating parameters

$Y \sim Categorical(\pi_1, \pi_2, \pi_3)$

$X_{Y=k} \sim \mathcal{N}(\mu_k, \sigma_k^2).$

$\mathcal{D} = \{x^{(i)}, y^{(i)}\}_{i=1}^{N}$

$$\begin{bmatrix} y_1^{(i)} \\ y_2^{(i)} \\ y_3^{(i)} \end{bmatrix}$$

$$\frac{\partial l}{\partial \pi_1} = 0$$

$$\frac{\partial l}{\partial \mu_3} = 0$$

$$f_N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(D \mid \theta)$$

$$= \quad '' \quad \prod_{i=1}^{N} p(x^i, y^i \mid \theta)$$

$$= \log \prod p(x^i, y^i \mid \theta)$$

$$= \sum_{i=1}^{N} \log p(x^i, y^i \mid \theta)$$

$$= \arg\max_{\theta} \sum \log p(y^{(i)} \mid \theta)\, p(x^i \mid y^i, \theta)$$

$$\sum_{i=1}^{N} \log \prod_{k=1}^{K} \underline{\pi_k}^{y_k^{(i)}} \underline{f_N(x^i \mid \mu_k, \sigma_k^2)}^{y_k^{(i)}}$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} y_k^{(i)} \log\left(\underline{\pi_k f_N(x^i \mid \mu_k, \sigma_k^2)}\right)}_{l(\theta \mid D)}$$

ML Applications of Bayes Rule

Bayes Rule

$$p(\, a \mid b \,) = \frac{p(\, b \mid a \,)\, p(a)}{p(b)} \qquad p(\, b \mid a \,) = \frac{p(\, a \mid b \,)\, p(b)}{p(a)}$$

$$p(\, \theta \mid \mathcal{D} \,) = \frac{p(\, \mathcal{D} \mid \theta \,)\, p(\theta)}{p(\mathcal{D})} \qquad p(\, \mathcal{D} \mid \theta \,) = \frac{p(\, \theta \mid \mathcal{D} \,)\, p(\mathcal{D})}{p(\theta)}$$

Poll 1

Which of these terms is the likelihood?

Select all that apply

$$\overset{A}{p(\theta \mid \mathcal{D})} = \frac{\overset{B}{p(\mathcal{D} \mid \theta)}\,\overset{C}{p(\theta)}}{\underset{D}{p(\mathcal{D})}} \qquad \overset{E}{p(\mathcal{D} \mid \theta)} = \frac{\overset{F}{p(\theta \mid \mathcal{D})}\,\overset{G}{p(\mathcal{D})}}{\underset{H}{p(\theta)}}$$

Bayes Rule

Terminology

Posterior Likelihood Prior

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)\, p(\theta)}{p(\mathcal{D})}$$

Two Applications of Bayes Rule

$$p(\, a \mid b \,) = \frac{p(\, b \mid a \,)\, p(a)}{p(b)}$$

$$p(\, \theta \mid \mathcal{D} \,) = \frac{p(\, \mathcal{D} \mid \theta \,)\, p(\theta)}{p(\mathcal{D})}$$

$$p(\, y \mid x \,) = \frac{p(\, x \mid y \,)\, p(y)}{p(x)}$$

MLE and MAP

Poll 2

Where do we plug in the pdf, e.g., $f(x) = \lambda e^{-\lambda x}$

$$\overset{A}{p(\theta \mid \mathcal{D})} = \frac{\overset{B}{p(\mathcal{D} \mid \theta)}\,\overset{C}{p(\theta)}}{\underset{D}{p(\mathcal{D})}}$$

MLE and MAP

Maximum likelihood estimation

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D} \mid \theta)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(y^{(i)} \mid \theta)$$

Maximum *a prosteriori* estimation

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta \mid \mathcal{D})$$

$$= \underset{\theta}{\operatorname{argmax}} \frac{p(\mathcal{D} \mid \theta) p(\theta)}{p(\mathcal{D})}$$

$$= \underset{\theta}{\operatorname{argmax}} \frac{\prod_{i=1}^{N} p(y^{(i)} | \theta) p(\theta)}{p(\mathcal{D})}$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{N} p(y^{(i)} \mid \theta)\, p(\theta)$$

Recipe for Estimation

MLE

1. Formulate the likelihood, $p(\mathcal{D} \mid \theta)$
2. Set objective $J(\theta)$ equal to negative log of the likelihood
   $$\mathrm{J}(\theta) = -\log p(\mathcal{D} \mid \theta)$$
3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

Recipe for Estimation

MAP

1. Formulate the likelihood times the prior, $p(\mathcal{D} \mid \theta)p(\theta)$
2. Set objective $J(\theta)$ equal to negative log of the likelihood times the prior
   $$\mathrm{J}(\theta) = -\log[p(\mathcal{D} \mid \theta)p(\theta)]$$
3. Compute derivative of objective, $\partial J / \partial \theta$
4. Find $\hat{\theta}$, either
   a. Set derivate equal to zero and solve for θ
   b. Use (stochastic) gradient descent to step towards better θ

Coin Flipping Example

Trick coin from pre-reading

Initially: no information about the coin, so we just default to a uniform belief about the Bernoulli parameter ϕ

Invoice: Weighted Coins — Customer: Torch Tricks, Inc.

Item	Quantity	Coin type, ϕ	$p(\phi)$
0% Heads Coin	40/200	0.0	0.20
20% Heads Coin	50/200	0.2	0.25
50% Heads Coin	80/200	0.5	0.40
80% Heads Coin	10/200	0.8	0.05
100% Heads Coin	20/200	1.0	0.10
	Total: 200		✓1.00

Poll 3

As we collect more and more data (more coin flips), will the peak of the likelihood curve increase or decrease?

A) Increase
B) Decrease
C) I have no idea

Coin Flipping Example

Trick coin from pre-reading

Suppose we discover information about the distribution of trick coin types?

How can we use this information both before and after flipping coins?

Invoice: Weighted Coins Customer: Torch Tricks, Inc.

Item	Quantity	Coin type, ϕ	$p(\phi)$
0% Heads Coin	40/200	0.0	0.20
20% Heads Coin	50/200	0.2	0.25
50% Heads Coin	80/200	0.5	0.40
80% Heads Coin	10/200	0.8	0.05
100% Heads Coin	20/200	1.0	0.10
	Total: 200		✓1.00

Coin Flipping Example

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\; p(\theta) \prod_{i=1}^{N} p(\, y^{(i)} \mid \theta \,)$$

Trick coin from pre-reading

Suppose we discover information about the distribution of trick coin types?

How can we use this information both before and after flipping coins?

$\mathcal{D}$	$p(\phi)\ \prod_{i=1}^{N} p(y^{(i)} \mid \phi)$
$\{\}$	$p(\phi)$
$\{H\}$	$p(\phi)\ \phi$
$\{H, T\}$	$p(\phi)\ \phi(1-\phi)$
$\{H, T, T\}$	$p(\phi)\ \phi(1-\phi)(1-\phi)$

Poll 5

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$$

posterior ∝ likelihood · prior

$$p(\theta \mid \mathcal{D}) \propto \prod p(y^{(i)}|\theta)\, p(\theta)$$

As the number of data points increases, which of the following are true?

Select ALL that apply

A. The MAP estimate approaches the MLE estimate
B. The posterior distribution approaches the prior distribution
C. The likelihood distribution approaches the prior distribution
D. The posterior distribution approaches the likelihood distribution
E. The likelihood has a lower impact on the posterior
F. The prior has a lower impact on the posterior

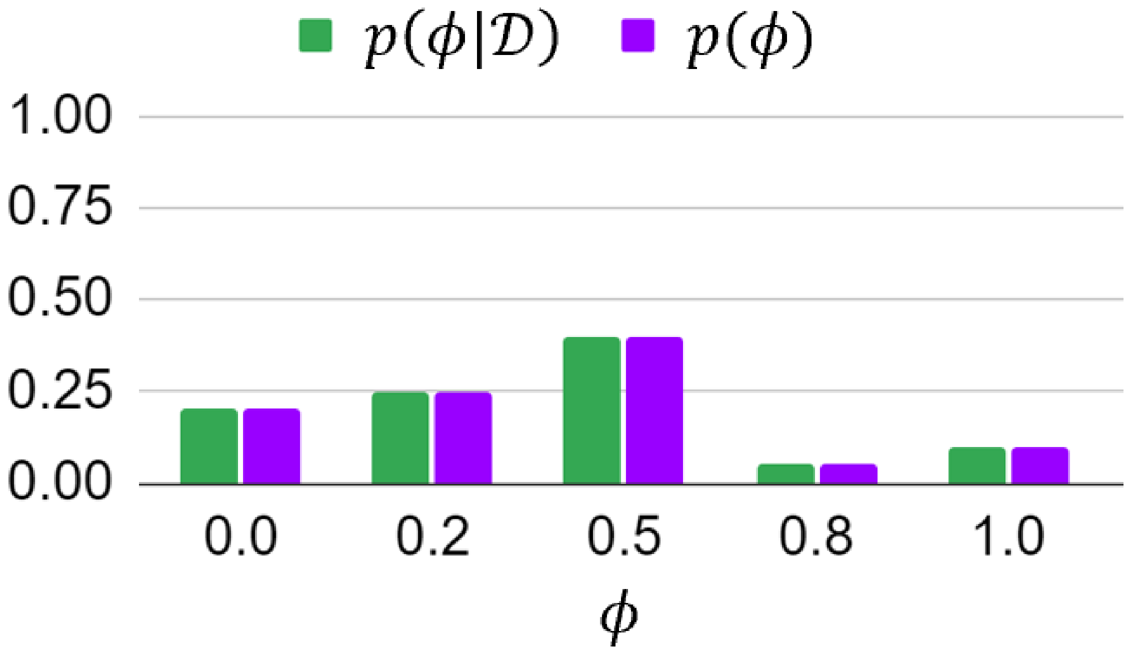
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p(y^{(i)} \mid \phi)\, p(\phi) \;=\; \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}\, p(\phi)$$

What happens as we flip more coins?

$N = 0$: $\mathcal{D} = \{\}$



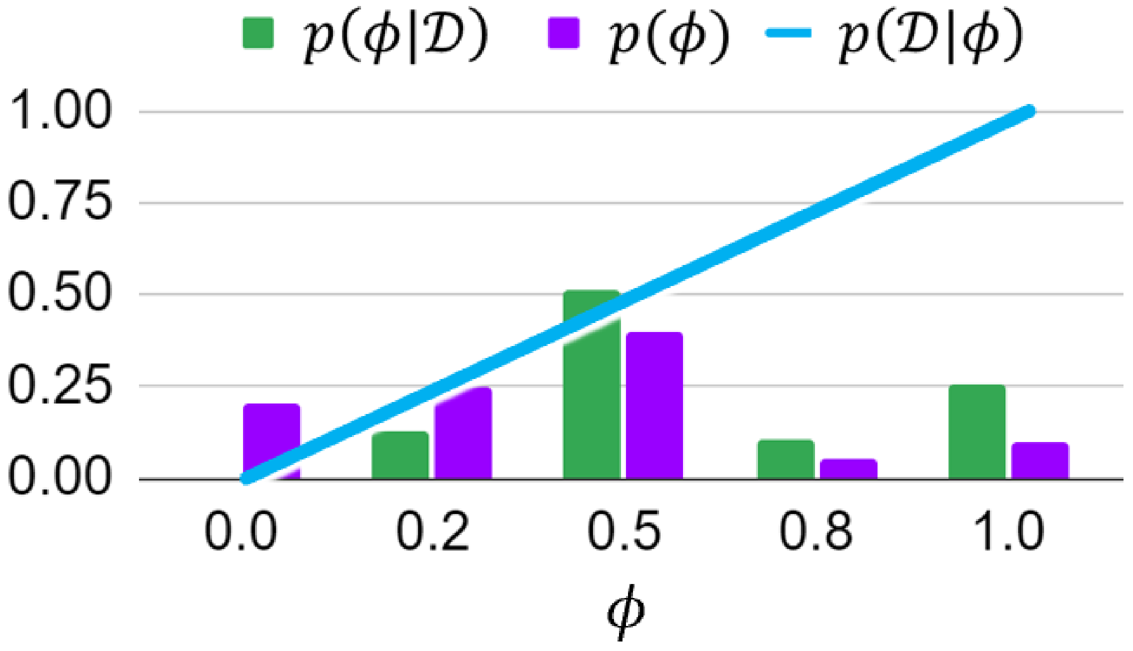
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p(y^{(i)} \mid \phi)\, p(\phi) \;=\; \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}\, p(\phi)$$

What happens as we flip more coins?

$N = 0$: $\quad \mathcal{D} = \{H\}$



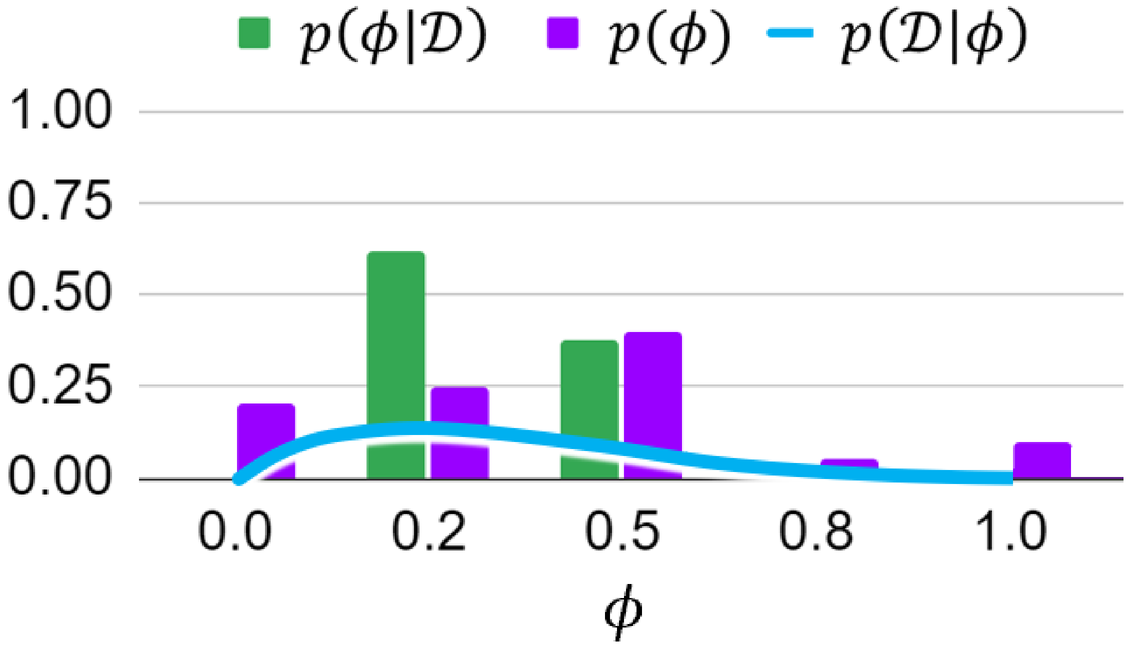
MAP as Data Increases

Given the ordered sequence of coin flip outcomes:

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p(y^{(i)} \mid \phi)\, p(\phi) \;=\; \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}\, p(\phi)$$

What happens as we flip more coins?

$N = 0$: $\mathcal{D} = \{H, T, T, T, T\}$



Prior Distributions for MAP

If the prior $p(\theta)$ is uniform, then MLE and MAP are the same!

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p(y^{(i)} \mid \phi)\, p(\phi) = \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}\, p(\phi)$$

https://www.desmos.com/calculator/kr7m2m6cf7

Prior Distributions for MAP

If the prior $p(\theta)$ is uniform, then MLE and MAP are the same!

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p(y^{(i)} \mid \phi)\, p(\phi) = \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}\, p(\phi)$$

Conjugate priors: when the prior and the posterior distributions are in the same family

Bernoulli likelihood with a Beta prior has Beta posterior

Categorical likelihood with a Dirichlet prior has Dirichlet posterior

Gaussian likelihood with a Gaussian prior has Gaussian posterior

https://www.desmos.com/calculator/kr7m2m6cf7

Prior Distributions for MAP

If the prior $p(\theta)$ is uniform, then MLE and MAP are the same!

$$\mathrm{p}(\mathcal{D} \mid \phi)\, p(\phi) = \prod_{i}^{N} p(y^{(i)} \mid \phi)\, p(\phi) = \phi^{N_{y=1}}(1-\phi)^{N_{y=0}}\, p(\phi)$$

Conjugate priors: when the prior and the posterior distributions are in the same family

Bernoulli likelihood with a Beta prior has Beta posterior

$$\phi^{N_{y=1}}(1-\phi)^{N_{y=0}}\, Beta(\alpha,\beta) = Beta(\alpha + N_{y=1}, \beta + N_{y=0})$$

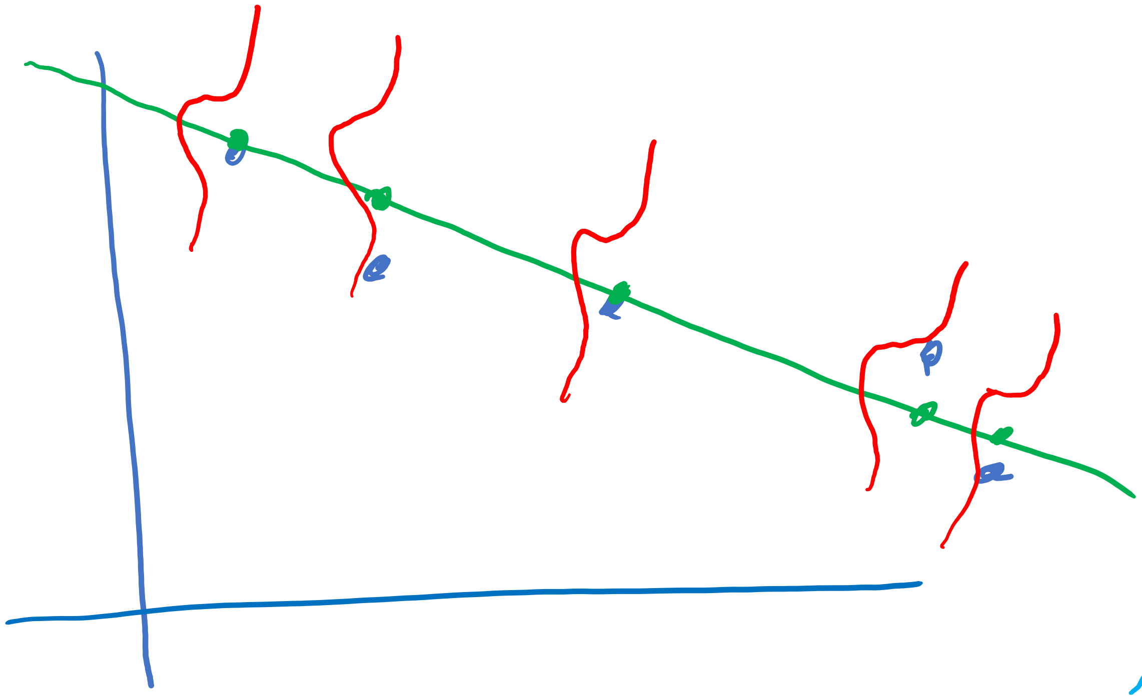
Tip: Think of the Beta distribution as having $\alpha - 1$ heads and $\beta - 1$ tails

https://www.desmos.com/calculator/kr7m2m6cf7

M(C)LE for Linear Regression

Probabilistic interpretation of linear regression

$$\hat{\theta}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i}^{N} p(\, y^{(i)} \mid \boldsymbol{x}^{(i)}, \boldsymbol{\theta}\,)$$



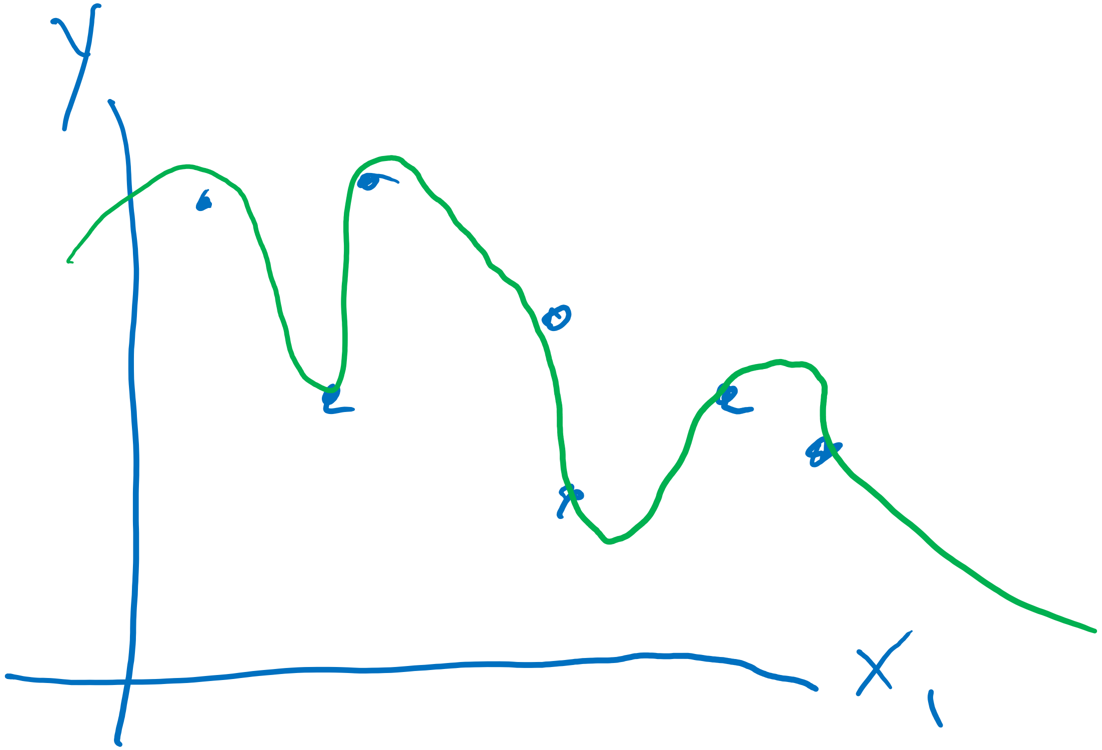
$$\begin{cases} y = w^T x + b + \epsilon \\ \epsilon \sim \mathcal{N}(0, \sigma) \end{cases}$$

$$y \sim \underbrace{\mathcal{N}(w^T x + b, \sigma)}_{\uparrow}$$

$$p(y \mid x, w, b) = \underline{\qquad\qquad}$$

MAP for Linear Regression

What assumptions are we making about our parameters?



$$\prod p(y^i \mid x^i, \vec{\theta})$$

$$\begin{bmatrix}1\\x_1\end{bmatrix} \xrightarrow{\phi} \begin{array}{cc}\begin{bmatrix}1\\x_1\\x_1^2\\x_1^3\\x_1^4\end{bmatrix} & \begin{array}{c}b\\w_1\\w_2\\w_3\\w_4\end{array}\end{array}$$

MAP for Linear Regression

Recall prereading example of Gausssian prior for Gaussian likelihood

$$p(\mu \mid \mathcal{D}) \propto p(\mu) \prod_{i=1}^{4} p(x^{(i)} \mid \mu)$$

$$= \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{1}{2\tau^2}(\mu-\nu)^2} \prod_{i=1}^{4} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x^{(i)}-\mu)^2}$$

Linear Regression with Gaussian prior on weights

Regularization and MAP

Linear Regression