$$Pr(X=k) = \binom{m}{k} p^k (1-p)^{m-k}$$

$p = Pr(X \geq k)$ = The prob under the null hypo that we would see at least this many flyers?

$$= \sum_{i=k}^{m} \binom{m}{k} \frac{1}{2}^k \frac{1}{2}^{m-k} = \frac{1}{2}^m \sum_{i=k}^{m} \binom{m}{k}$$

Ex. - $m=10, k=9, p=0.01$
$m=10, k=5 \quad p=0.62$
- $m=100, k=90, p \approx 10^{-6}$
$m=100, k=50, p=0.54$

Reject H_0 if evidence unlikely under H_0 distribution

Real error

	D_1	D_2	. . .	D_m
a_1	0.1	0.05	. .	0.05
a_2	0.2	0.03	. .	0.07
d_i	+0.1	-0.02	- - -	+0.02

d_i : $\frac{error_2}{error_1}$

T is our RV, = normalized average. differences in error.

$$\bar{d} = \frac{1}{m} \sum_{i=1}^{n} d_i \qquad S_d = \sqrt{\frac{1}{m-1} \sum_{i=1}^{m} (d_i - \bar{d})^2}$$

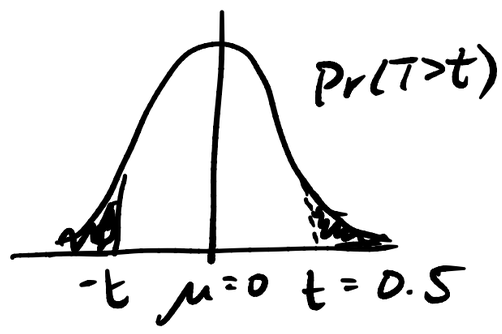
$$t = \frac{\bar{d} - \mu}{S_d / \sqrt{m}}$$

Assume T follows a Student t-distribution

μ = expect average diff.

H_0: $\mu = 0$

H_1: $\mu > 0$ (a_2 better than a_1)



$p = Pr(T > t)$ if p is small, the evidence is unlikely under H_0, so reject H_0

$p = Pr(T > t)$ or $Pr(T < -t)$

NOV. 26, 2019: find (latent) features $h \in \mathbb{R}^k$ for $x \in \mathbb{R}^d$ that explain x and have certain properties

$k < d$ → goal: lower-dimensional representation, compactness, avoiding overfitting

(1) Learn dictionary $D \in \mathbb{R}^{k \times d}$ $\quad x \approx h D = h_1 \vec{d}_1 + h_2 \vec{d}_2 + \ldots + h_k \vec{d}_k$ (h: $1 \times k$, D: $k \times d$; $h_1, \ldots, h_k$ scalars)

$$\min_{D, h_1, \ldots, h_n} \sum_{i=1}^{n} \| x_i - h_i D \|_2^2$$

(x_i: $1 \times d$, h_i: $1 \times k$, D: $k \times d$; $D, h_1, \ldots, h_n$ vectors)

Found D and $h_1, \ldots, h_n$ $\quad \phi(x_i) = h_i$

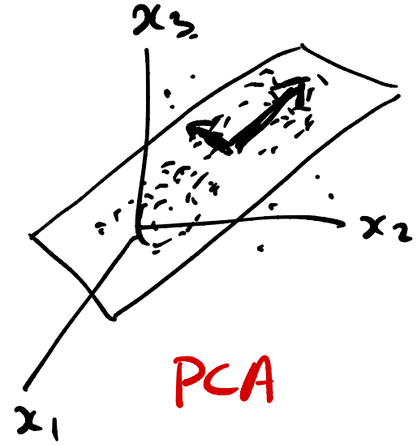
For new x, $\phi(x) = \arg\min_h \| x - h D \|_2^2$

Solution to (1): Let $X = U \Sigma V^T$ ($n \times d$; V^T: $d \times d$)

$D = V^T_{:,1:k}$ ($k \times d$) $\quad H = \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix} = U_{:,1:k} \Sigma_k$

$h_i = U_{i,1:k} \Sigma_k$

$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_d \end{bmatrix}$ $\quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0$



How is this different from NNs? Equivalent for one setting

Linear auto-encoder $\quad x \in \mathbb{R}^d \to h \in \mathbb{R}^k \to \ldots \to x$ ($W^{(1)}$, $W^{(2)}$)

Produce $\hat{x} \approx x$. $\quad$ linear activations $\quad \hat{x} = x W^{(2)} W^{(1)}$

$$\min_{W^{(1)}, W^{(2)}} \sum_{i=1}^{n} \| \underbrace{(x_i W^{(2)}) W^{(1)}}_{\hat{x}_i} - x_i \|_2^2 \rightarrow W^{(2)} = V_{:,1:k}$$

$$W^{(1)} = V_{:,1:k}^T$$

$$h_i = x_i W^{(2)} = U_{i,:} \Sigma V^T V_{:,1:k} = U_{i,1:k} \Sigma_k$$

Otherwise typically different, why would we use factorization?

Let's encode specific properties: high-dimensional h that is sparse

h is sparse if a small # of elements are non-zero, $h_i \in \mathbb{R}^k$, $k >> d$.

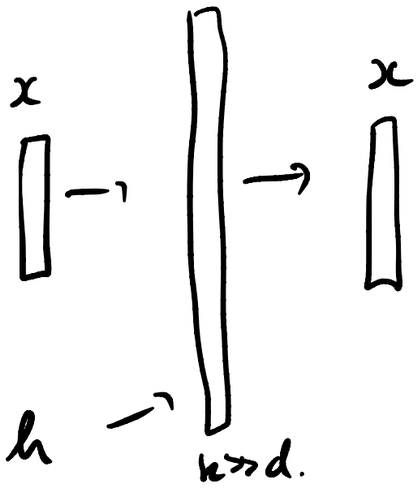
$$\min_{D, h_1, \ldots, h_n} \sum_{i=1}^{n} \| x_i - \underbrace{h_i}_{1\times k} \underbrace{D}_{k \times d} \|_2^2 + \lambda \sum_{i=1}^{n} \|h_i\|_1$$

$\hat{x}_i = x_i$

$$+ \lambda_0 \sum_{j=1}^{k} \|d_j\|_2^2$$

Sparse coding

NN:



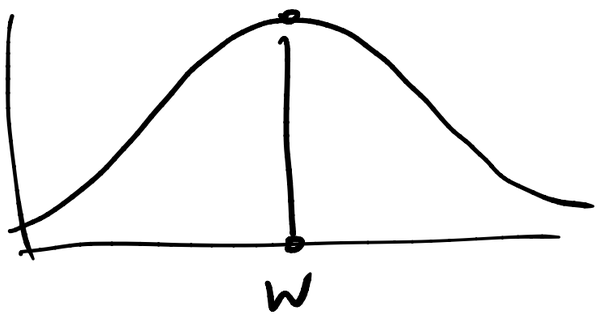
$$\sum \| x_i W^{(2)} W^{(1)} - x_i \|_2^2 + \lambda \| x_i W^{(2)} \|_0$$

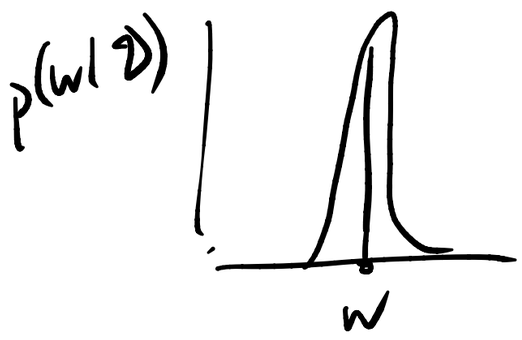
Nov. 28, 2019: Bayesian linear regression

$p(y|x) = \mathcal{N}(\mu = x^T w, \sigma^2)$ $\quad \mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$

$p(w) = \mathcal{N}(0, \lambda^{-1} I)$

MAP: $\operatorname{argmax}_w p(w|\mathcal{D})$ Instead, if we've Bayesian maintain $p(w|\mathcal{D})$





$$p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)\,p(w)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|w)\,p(w)}{\int p(\underbrace{\mathcal{D}|w')\,p(w')}_{p(\mathcal{D},w')}\,dw'}$$

$p(w|\mathcal{D}) = \mathcal{N}(w|m, S)$ $\quad m = (X^T X + \lambda\sigma^2 I)^{-1} X^T y$

$$S = \sigma^2 (X^T X + \lambda I)^{-1}$$

Conjugate prior: prior for w s.t that the is of the same form with different parameters

$f(x, w) = x^T w$

$\bar{f}(x) = E[f(x, \mathbf{w})] = \int_{\mathbb{R}^d} f(x, w)\, p(w|\mathcal{D})\, dw = m^T x$ (w: random variable)

$Var(f(x, w)) = \int (f(x, w) - E[f(x, w)])^2 p(w|\mathcal{D})\, dw$

$= x^T S x$

If we had picked $p(w)$ = Laplace distr., $p(w|\mathcal{D})$?

- $p(y)$ $y \in \{0, 1\}$, parameter $w \in [0, 1]$ (Bernoulli parameter)

Bernoulli

$\alpha - 1$ successes
$\beta - 1$ failures

Beta distribution $\text{Beta}(\alpha, \beta)$

$$p(w|\alpha, \beta) = \frac{w^{\alpha-1}(1-w)^{\beta-1}}{\text{normalization term}}$$

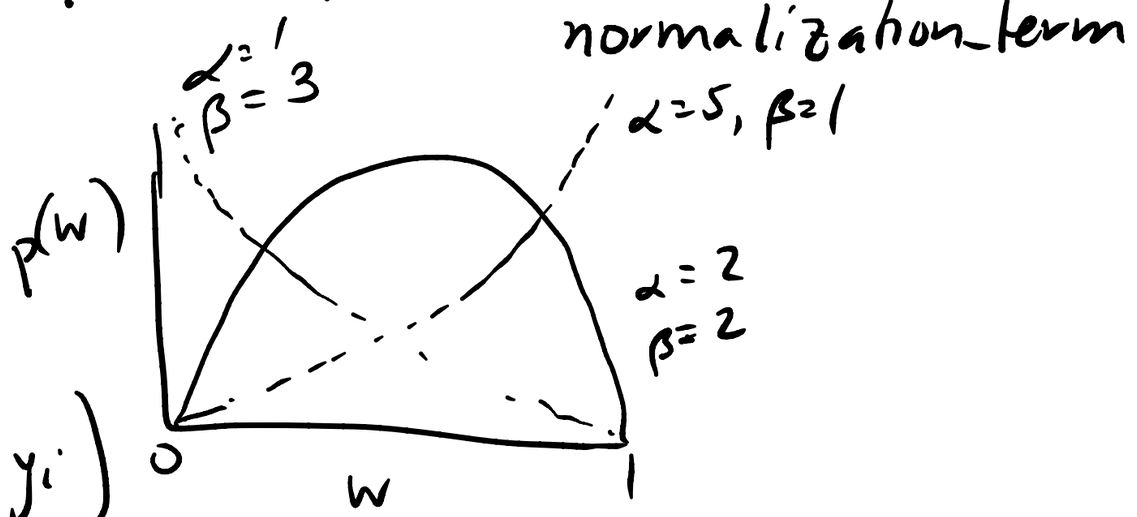
$\mathcal{D} = \{y_i\}_{i=1}^n \quad y_i \in \{0, 1\}$

$p(y)$ Bernoulli (w)

$p(w)$ $\text{Beta}(\alpha, \beta)$

$$p(w|\mathcal{D}) = \text{Beta}\left(\alpha + \sum_{i=1}^n y_i, \beta + n - \sum_{i=1}^n y_i\right)$$

α' ← # successes; β' ← # failures



$$p(y|\mathcal{D}) = \int p(y|w)\, p(w|\mathcal{D})\, dw$$ Posterior predictive distribution

Exercise: What do you do with new data?

$\{y_i\}_{i=n+1}^{n+m}$ e.g. how does $p(w|\mathcal{D})$ change for Bernoulli + Beta?

$$p(w|\mathcal{D}) = \text{Beta}\left(\alpha + \sum_{i=1}^{n+m} y_i,\ \beta + n + m - \sum_{i=1}^{n+m} y_i\right)$$

$$\alpha' + \sum_{i=n+1}^{n+m} y_i,\ \beta' + m - \sum_{i=n+1}^{n+m} y_i$$

$$p_n(w) \doteq p(w|\mathcal{D}_n) = \text{Beta}(\alpha', \beta')$$

$$p_n(w|\mathcal{D}) = \text{Beta}\left(\alpha' + \sum_{i=n+1}^{n+m} y_i,\ \beta' + m - \sum_{i=n+1}^{n+m} y_i\right)$$