1 Lecture 1

KNN

Given data $D = \{x^i, y^i\}$, KNN predicts on a new x the label:

$$y^i \text{ where } \operatorname{argmin}_{i \in \{1 \dots N\}} \|x - x^i\| \text{ is minimal}$$

1.1 Problems

- The storage cost involved in storing all x^i.
- The computation cost involved in running through all x^i.
- The existence of all outliers.
  - Now additionally suppose that this is in several dimensions.

2 Lecture 2

Linear Regression

Remark 2.1. We always assume that D represents a set of training data $\{x^i, y^i\}$.

Definition 2.2. Given D, OLS finds the $\mathbf{w_1}$ and w_2 such that if $\mathbf{w} = \begin{bmatrix} \mathbf{w_1} \\ w_2 \end{bmatrix}$, then $\|Xw - y\|_2$ is minimized.

Note that this X has been augmented by a column of 1s. That is, $X = \begin{bmatrix} X_{\text{original}} & 1 \\ \vdots & 1 \\ \vdots & \vdots \\ & 1 \end{bmatrix}$. For this reason, the dimension of X is $n \times (d+1)$, $\mathbf{w}$ is $d+1$ and y is $n \times 1$.

Example 2.3. Suppose that for some D, all x^i coincide. What are the best values for $\mathbf{w_1}$ and w_2 then?

See that $\|Xw - y\|_2$ is also, if squared,

$$\sum_{j=1}^{n} (x^j \mathbf{w_1} + w_2 - y^j)^2$$

$$= \sum_{j=1}^{n} (w_2 - y^j)^2$$

Since $x^j \mathbf{w_1}$ is always a constant

$$\implies \text{ optimal } w_2 = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y^i$$

Theorem 2.4. The optimal $\mathbf{w}$ is given by $\hat{\mathbf{w}} = X^T X^{-1} X^T y$.

Proof.

$$1/2(\|Xw-b\|_2)^2 \\ = 1/2(Xw-b)^T(Xw-b)$$

Now set the derivative to 0

$$0 = \frac{\partial(Xw-b)}{\partial w}1/2\frac{\partial((Xw-b)^T(Xw-b))}{\partial(Xw-b)}$$
$$0 = X^T1/2(2)(Xw-b)$$
$$0 = X^T(X\hat{w}-b)$$

And the result comes easily

□

Remark 2.5. X^TX^{-1} does not always exist.

Theorem 2.6. X^TX is invertible if and only if Rank$X = d+1$ and $d+1 \leq n$.

Proof.

$\Longrightarrow$

Then $d+1 = \text{Rank}X^TX \leq \text{Rank}X \leq \min\{n, d+1\}$. This must mean that $\min\{n, d+1\} = d+1$, meaning that Rank$X = d+1$ and $d+1 \leq n$.

$\Longleftarrow$

Suppose that there exists u such that $X^TXu = 0$. Then $u \cdot X^TXu = u \cdot 0$, which means that $(Xu, Xu) = 0 \implies Xu = 0$. This is a contradiction, however, since X has rank $d+1$. We conclude that X^TX is invertible. □

Remark 2.7. Suppose that the inverse does not exist. Then either we use the pseudo-inverse (see your 357 notes for more information). The pseudoinverse can be shown to be the unique existing vector that, among those vectors minimizing $\|Xw-b\|_2^2$, is minimal in norm.

Using the pseudoinverse, the new best estimate for $\mathbf{w}$ works out to be $(X^TX)^+X^Ty = X^+y$.

Remark 2.8. The other estimate for $\mathbf{w}$ is obtained using ridge regression

$$\text{argmin}_{w\in\mathbb{R}^{d+1}}\|X\mathbf{w}-y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$$

$\mathbf{w} = (X^TX+\lambda I)^{-1}X^Ty$

2.0.1 Geometry and Probability

Remark 2.9. Note that since $X^T(X\hat{w} - y) = \vec{0}$, we find that

$X\hat{w} - y$ is orthogonal to each column of X and, hence, to the span of the columns in X.

Remark 2.10. The Gaussian probability model makes the assumption that

$$\mathbb{P}(y^i|x^i) = N(\bar{\mathbf{w}}^T x^i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(y^i - \mathbf{w}^T x^i)^2\right)$$

We wish to choose $\bar{\mathbf{w}}$ so that

$$\mathbb{P}\left(\bigcap_{i=1}^{n}(y^i, x^i)\right)$$

is maximized. If we assume that all points in D are iid, then this is equivalent to finding a value for $\mathbf{w}$ maximizing

$$\text{argmax}_{\mathbf{w}\in\mathbb{R}^{d+1}} \prod_{i=1}^{n} \mathbb{P}(y^i|x^i) = \text{argmin}_{\mathbf{w}\in\mathbb{R}^{d+1}} \sum_{i=1}^{n} \frac{1}{2\sigma^2}(y^i - \mathbf{w}^T x^i)^2 = \text{argmin}_{\mathbf{w}\in\mathbb{R}^{d+1}} \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{1}{2}(y^i - \mathbf{w}^T x^i)^2$$

Remark 2.11. In general, the loss function

$$\frac{1}{N}\sum_{i=1}^{n}(y^i - \mathbf{w}^T x^i)^2$$

can be generalized to be

$$\frac{1}{N}\sum_{i=1}^{n} l_2(y^i, \mathbf{w}^T x^i) \qquad (\alpha)$$

where $l_2(y, \hat{y}) = (y - \hat{y})^2$. In general, we can use any loss function in place of l_2, and so α can be expressed as

$$\frac{1}{N}\sum_{i=1}^{n} \text{argmin}_f\, l(y^i, f(x^i)) \qquad (\beta)$$

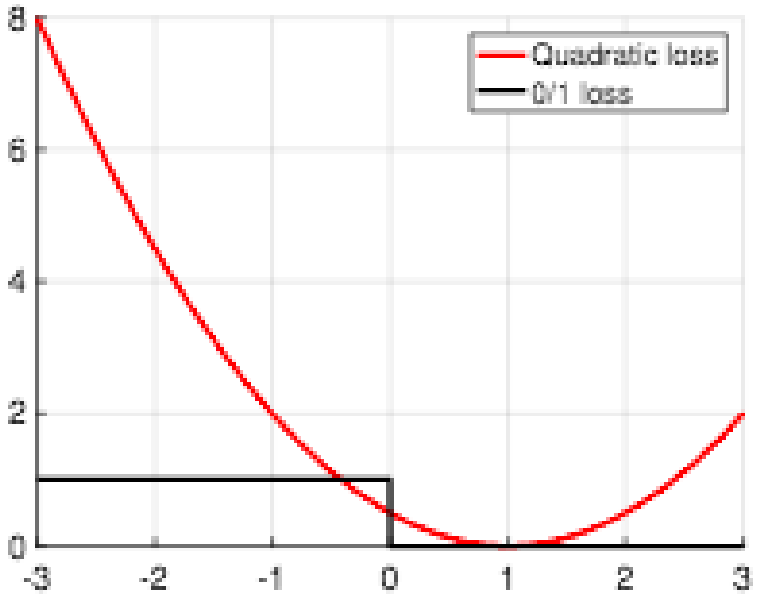
is generally known as the ERM minimization algorithm.

Remark 2.12. Suppose that we wish to use binary classification with linear regression. That is given D with labels for y in $\{\pm 1\}$, we find the optimal w such that $\hat{y} = sign(\mathbf{w}^T x)$ is minimized under square loss.

Note that the loss function admits simplification here:

$$\frac{1}{2}(y - \hat{y})^2 = \frac{1}{2}(y^2 - 2y\hat{y} + (\hat{y})^2) = (1 - y\hat{y}) = (1 - y\hat{y})^2/2$$

If we compare this to the binary loss function $1_{y\neq\hat{y}} = 1[y\hat{y} \leq 0]$, then we see in the following graph (x axis traces the value of $y\hat{y}$ and y axis traces the error), that the square loss function penalizes the label classification for being correct.



Question 2.13. Unresolved: What are some justifications for least squares? Why might we use ridge regression?

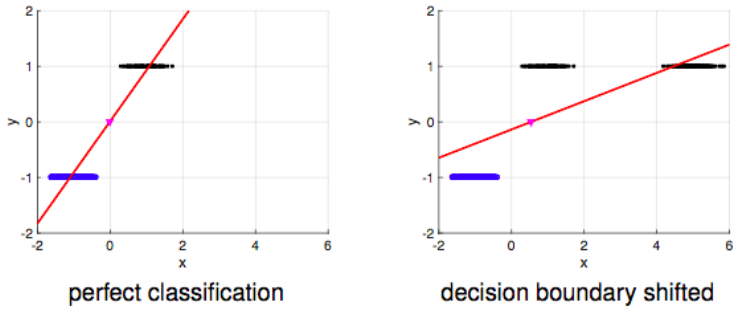
3 Lecture 3 – Logistic Regression

Remark 3.1. Recall that linear regression for classification finds the optimal $\mathbf{w}$ for the regression problem over D where labels for $\vec{y}$ are either 1 or -1. Ultimately, we are trying to minimize

$$\sum_{i=1}^{N}(y^i - \mathbf{w}^T x^i)^2$$

which is why the location of decision boundary does not reflect a good choice of separation but the location where a line minimizing square loss just happened to pass through the x-axis. This creates figures like the one following:

In our case:

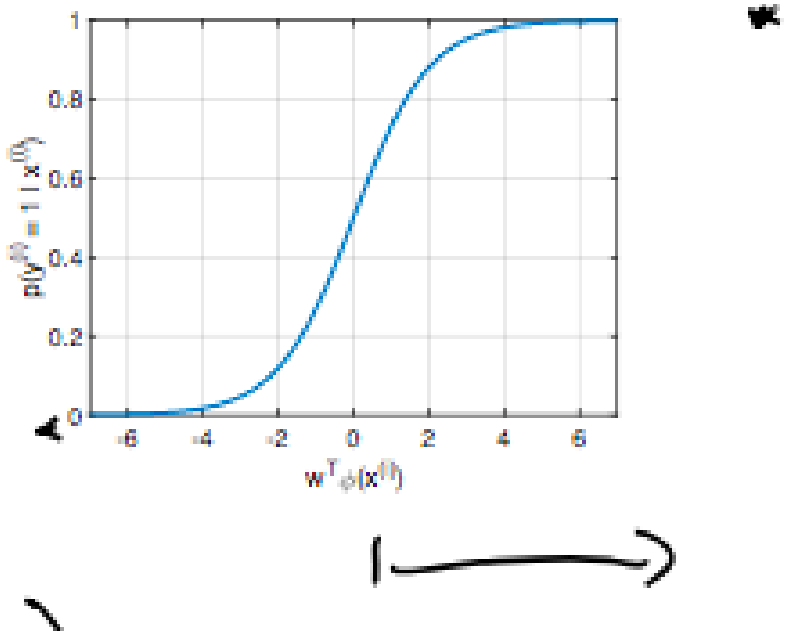


In some sense, points that are easy to classify are penalized for being too correct.

Definition 3.2. In logistic regression, we define a probability function for an instance:

$$\mathbb{P}(y = 1|x) = \frac{1}{1 + \exp\left(-\mathbf{w}^T\phi(x)\right)} \qquad (\alpha)$$

It satisfies the nice property that as saturation increases (greater in absolute value) the derivative of likelihood decreases. It will also correct the problem we observed earlier by not penalizing points that are very correct.



Definition 3.3. The logistic regression probability function (for a single point) for the purpose of binary classification is

$$\mathbb{P}(y = y^i|x) = \frac{1}{1 + \exp\left(-y^i \mathbf{w}^T \phi(x^i)\right)}$$

Theorem 3.4. This equation is consistent with α.

Proof.

$$\mathbb{P}(y = -1|x) = 1 - \mathbb{P}(y = 1|x) = 1 - \frac{1}{1 + \exp\left(-\mathbf{w}^T \phi(x)\right)} = \frac{1}{1 + \exp\left(\mathbf{w}^T \phi(x)\right)}$$

□

Theorem 3.5. $\text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + \exp\left(-y^i \mathbf{w}^T \phi(x^i)\right))$ is the loss function that assuming iid points in D, we want must minimize under logistic regression.

Proof. Observe, first, that if there is a $\mathbf{w}$ that satisfies

$$\text{argmax}_{\mathbf{w}} \prod_{i=1}^{n} \mathbb{P}(y = y^i|x^i)$$

then this same w satisfies

$$\text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} -\log(\mathbb{P}(y = y^i|x^i))$$

Now, if we substitute our expression for $\mathbb{P}(y = y^i|x^i)$ from above, we have

$$= \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + \exp\left(-y^i \mathbf{w}^T \phi(x^i)\right))$$

□

Remark 3.6.

Comparison

Linear regression

Program:

$$\min_{\boldsymbol{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \frac{1}{2} (1 - y^{(i)} \underbrace{\boldsymbol{w}^T \phi(x^{(i)})}_{F(x^{(i)}, w)})^2$$

$F(x^{(i)}, w, y^{(i)})$

Logistic regression

Program:

$$\min_{\boldsymbol{w}} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \log\left(1 + \exp(-y^{(i)} \underbrace{\boldsymbol{w}^T \phi(x^{(i)})}_{F(x^{(i)}, w)})\right)$$

$F(x^{(i)}, w, y^{(i)})$

Note that both the loss function for linear regression and logistic regression use this term $-y^i \mathbf{w}^T \phi(x^i)$ which we denote as $F(x^i, w, y^i)$. One loss function is

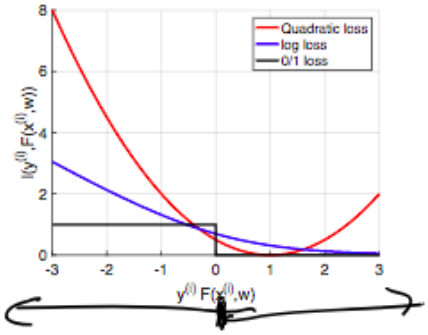
$$\frac{1}{N} \sum_{i=1}^{N} (1 + F(x^i, w, y^i))^2.$$

while the other is

$$\frac{1}{N}\sum_{i=1}^{N}\log(1+\exp(F(x^i,w,y^i)))$$

Indeed, both are (again) representatives of the ERM algorithm. If we denote $F(x^i,w)=\mathbf{w}^T\phi(x^i)$, then we can represent the ERM algorithm as choosing $\mathbf{w}$ to minimize

$$\frac{1}{N}\sum_{i=1}^{N}l(y^i,F(x^i,w))$$



$y=+1$
$\hat{y}=$ Large
Log loss small

We see in the scenario above, where $\hat{y}$ is large and y is 1, that the newfound logistic loss function does not penalize us for extreme correctness.

Remark 3.7. Since the minimum $\mathbf{w}$ in the logistic loss function admits no close form solution, we have to employ optimization. We can, for example, employ gradient descent.

To solve

$$\min_{\boldsymbol{w}} f(\boldsymbol{w}) := \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \log\left(1+\exp(-y^{(i)}\underbrace{\boldsymbol{w}^T\phi(x^{(i)})}_{F(x^{(i)},w)})\right)$$

we can use its gradient:

$$\nabla_{\boldsymbol{w}} f(\boldsymbol{w}) = \sum_{(x^{(i)},y^{(i)})\in\mathcal{D}} \frac{-y^{(i)}\phi(x^{(i)})\exp(-y^{(i)}\boldsymbol{w}^T\phi(x^{(i)}))}{1+\exp(-y^{(i)}\boldsymbol{w}^T\phi(x^{(i)}))}$$

Simple algorithm: Initialize $t=0$, $\boldsymbol{w}_t$, and stepsize α

- Compute gradient $\mathbf{g}_t=\nabla_{\boldsymbol{w}} f(\boldsymbol{w}_t)$
- Update parameters $\boldsymbol{w}_{t+1}\leftarrow\boldsymbol{w}_t-\alpha\mathbf{g}_t$
- Update $t\leftarrow t+1$

Unresolved: At some point, look into how matrix algebra can be used to get the gradient as we have it above.

`https://en.wikipedia.org/wiki/Matrix_calculus#Relation_to_other_derivatives`

4 Lecture 4 – Optimization Basics

Definition 4.1. A set S is convex if, for any two points it contains, it contains the line segment between the points. Formally, if x and y are points, then

$$\{\alpha x + (1-\alpha)y : 0 \leq \alpha \leq 1\} \subseteq S$$

Definition 4.2. The convex hull of a set S (not necessarily convex) is the smallest convex superset containing S. It is the intersection of all convex sets containing S. Given a finite set $S = \{x_1 \dots x_k\}$,

$$\{\sum_{i=1}^{k} \alpha_i x_i | \alpha_i \geq 0\}$$

is also the convex hull.

Remark 4.3. We will use ,throughout this note, the fact that convexity is preserved under intersection.

Example 4.4. In optimization, people define polyhedron to be a set of the form, for some A and b,

$$\{x|Ax \leq b\} \tag{α}$$

The resulting figure, when viewed in a multidimensional plot does in fact resemble what geometers think of as a polyhedron. See wikipedia for an image of the geometer's conception of a polyhederon. Proving that a polyhedron is convex can be done by proving that the following sets:

$$\{x|a^T x \leq b\} \text{ and } \{x|a^T x = b\}$$

are convex. In this case, note that b is a scalar. Then, after appealing to the fact that an intersection of convex sets is also convex, notice that (α) is really those x such that $a_i^T x \leq b_i$ where a_i is the ith row of A and b_i the ith entry of b.

Example 4.5. The epigraph of a convex function f is the set

$$\{(x, r)|f(x) \leq r\}$$

It is a simple exercise to show that the epigraph is a convex set if and only if f is a convex function.

Definition 4.6. The standard definition of convexity is that where $0 \leq \alpha \leq 1$, for any $x, y \in \mathcal{D}(f)$, the domain of f, we have that

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

Proposition 4.7. Every norm is convex.

Proof.

$$\|\alpha(x) + (1-\alpha)y\| \leq \|\alpha x\| + \|(1-\alpha)y\| = \alpha\|x\| + (1-\alpha)\|y\|$$

□

Remark 4.8. Convexity in 1 dimension can be shown to be equivalent to the conditions that f' be monotonically non-decreasing and f'' be non-negative (replace "non" with "strictly") to get strict convexity. There are analogs in the multidimensional dimensional case for convexity and λ-strong convexity (note that $\lambda > 0$)

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\lambda\alpha(1-\alpha)}{2}\|x-y\|^2$$

Ordinarily we would have, as an extension of monotonically non-decreasing first derivative,

$$f(y) = f(x) + \nabla f(x)^T(y-x)$$

With $\lambda-$strong convexity, we now have

$$f(y) = f(x) + \nabla f(x)^T(y-x) + \frac{\lambda}{2}\|y-x\|^2$$

and

$$\nabla^2 f(x) \geq 0$$

is replaced by, for $\lambda-$ strong convexity,

$$\nabla^2 f(x) \geq \lambda I.$$

See `http://www.princeton.edu/~amirali/Public/Teaching/ORF523/S16/ORF523_S16_Lec7_gh.pdf` for proofs in the multivariate case.

Remark 4.9. Assuming that $f : \mathbb{R}^d \to \mathbb{R}$, note that the requirements

$$\nabla^2 f(x) \geq 0$$

and

$$\nabla^2 f(x) \geq \lambda I$$

are biconditional with the statement in the first that $\nabla^2 f(x)$ have non-negative eigenvalues and, in the second, that $\nabla^2 f(x)$ have eigenvalues greater than λ. For since $\nabla^2 f$ is symmetric, it admits a diagonalization with D, consisting of its eigenvalues along the diagonal. Let M represent the hessian from now on. Then $(Px)^T DPx \geq 0$ if and only if $x^T Mx \geq 0$ (note that here $P^T DP = P^{-1}DP = M$). But $(Px)^T DPx = y^T Dy$, where y assumes all values in $\mathbb{R}^d$.

Unresolved 4.10. Why is $P^{-1} = P^T$ in this case?

Proposition 4.11.

- Linear combinations of convex functions are convex, where the weights are positive. This can be easily shown using the standard definition of convex.
- The supremum of convex functions is convex.
  - This proof is not known.

- The application of convex f to $Aw+b$ where A and b are any matrix and vector is convex.
  - This proof is not known.

Remark 4.12. Recall that ERM applied to logistic regression and linear regression are generalized by $\sum_{j=1}^{n} l(y^i w^T x^i)$. Applying the preceding proposition to this generalization, we find that the loss function ERM seeks to minimize is convex.

Remark 4.13. In the case that a function is not differentiable, say $f(x) = |x|$, then we can define the sub-differential, which is

$$\delta f(x) = \{s | \forall y, f(y) \geq f(x) + s^T(y-x)\}$$

In the case that f is differentiable at x, then there is only s whose value is $\nabla f(x)$.

Unresolved 4.14. Why is this true?

Remark 4.15. If 0 is in the subdifferential – that is, suppose that $0 \in \delta(y)$ – then $y = \inf_x f(x)$.

Theorem 4.16. Jensen's inequality says that if $f : \mathbb{R}^d \to \mathbb{R}$ is convex then

$$\mathbb{E} f(X) \geq f(\mathbb{E} X)$$

Here, we assume that X is a random variable.

Proof.

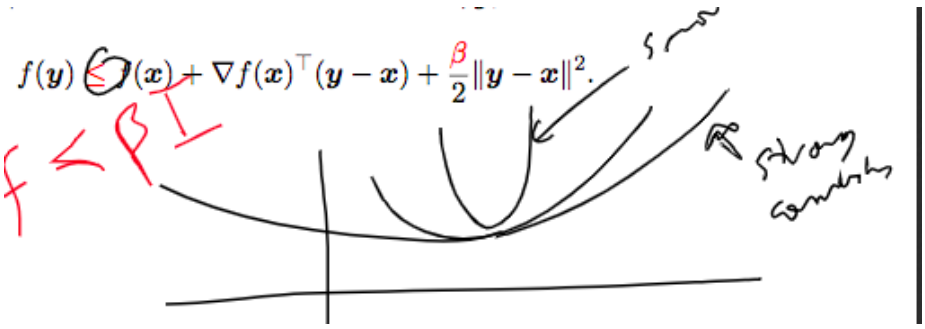
$$\mathbb{E} f(\boldsymbol{X}) \geq \mathbb{E}\left(f(\boldsymbol{y}) + \boldsymbol{s}^\top(\boldsymbol{X}-\boldsymbol{y})\right) = f(\boldsymbol{y}) + \boldsymbol{s}^\top \mathbb{E}(\boldsymbol{X}-\boldsymbol{y}) = f(\boldsymbol{y}).$$

□

Definition 4.17. A function is β–smooth when the reverse inequality (with respect to λ-strong convexity) holds:

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{\beta}{2}\|y-x\|^2$$

Remark 4.18. If a function is both β smooth and λ-strong convex, then it is well behaved in the sense that it is bounded by two quadratic functions.



Hi, I am r^2 and b^2.